C 13.22/sec. B: 74B/4

Public Documents Division
Oklahoma Department of Libraries

APR 19 1971

SECTION B

A UNITED STATES DEPARTMENT OF COMMERCE PUBLICATION

JOURNAL OF RESEARCH

of the NATIONAL BUREAU OF STANDARDS

Volume 74B, Number 4 Oct.-Dec. 1970

Mathematical Sciences

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards

Journal of Research

of the

National Bureau of Standards

B. Mathematical Sciences

Oct.–Dec. • 1970

VOLUME • 74B

NUMBER • 4

Editor: Morris Newman

Associate Editors: F.W. Olver and J. R. Edmonds

Publication dates: Mar. 10, June 10, Sept. 10, Dec. 10, 1970

J. Res. Nat. Bur. Stand. (U.S.), B CODEN: JNBBA 74B(4) 231–332 (1970)

JOURNAL OF RESEARCH

The National Bureau of Standards Journal of Research reports research and development in the fields of activity shown at right. Also included from time to time are survey articles on topics closely related to the Bureau's scientific and technical program.

The Journal is published in three separate sections as follows:

A. Physics and Chemistry

Contains papers of interest primarily to scientists working in these fields. Issued six times a year. Annual subscription: domestic, $9.50; foreign, $11.75. Single copy, varies.

B. Mathematical Sciences

Presents studies and compilations designed mainly for the mathematician and the theoretical physicist. Issued quarterly. Annual subscription: domestic, $5.00; foreign, $6.25. Single copy $1.25.

C. Engineering and Instrumentation

Reports research and development results of interest chiefly to the engineer and the applied scientist. Issued quarterly. Annual subscription: domestic, $5.00; foreign, $6.25. Single copy $1.25.

U.S. DEPARTMENT OF COMMERCE
Maurice H. Stans, Secretary
NATIONAL BUREAU OF STANDARDS
Lewis M. Branscomb, Director

Order all publications from the Superintendent of Documents
U.S. Government Printing Office, Washington, D.C. 20402.

Use of funds for printing this publication approved by the Director of the Bureau of the Budget (June 22, 1966).

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

A Property of the Triangle Groups

Joseph Lehner*

**Institute for Basic Standards, National Bureau of Standards,
Washington, D.C. 20234**

(August 27, 1970)

The F-groups are the groups possessing faithful representations by Fuchsian groups of the first kind; their presentations are known explicitly. Among the F-groups are the well-known triangle groups $G=\{x, y|x^p=y^q=(xy)^r=1\}$. If p, q, r are distinct prime integers, every proper normal subgroup of finite index in G has no elements of finite order. In this paper it is proved that among the F-groups only the triangle groups with distinct prime p, q, r have this property.

Key words: Element of finite order; F-group; normal subgroup; triangle group.

1. The F-groups of J. Nielsen consist of those groups that can be faithfully represented by Fuchsian groups of the first kind. Such groups are known to be exactly those that have presentations of the form

$$\left\{e_1, \ldots, e_s, p_1, \ldots, p_t, a_1, b_1, \ldots, a_g, b_g \mid e_1^{l_1}=\ldots=e_s^{l_s} = \prod_{i=1}^{s} e_i \prod_{j=1}^{t} p_j \prod_{k=1}^{g} a_k b_k a_k^{-1} b_k^{-1}=1\right\} \tag{1.1}$$

where $s, t, g \geq 0$, $l_i \geq 2$ and

$$g-1+\frac{1}{2}\left(t+\sum_{i=1}^{s}\left(1-\frac{1}{l_i}\right)\right)>0. \tag{1.2}$$

For each such group G it has been proved that there exists a proper normal subgroup of finite index with no elements of finite order [1, 2].[1]

On the other hand, the particular F-group

$$\Delta=\{e_1, e_2, e_3|e_1^2=e_2^3=e_3^7=e_1e_2e_3=1\}$$

has the property that *every* proper normal subgroup of finite index is torsion free.

We shall say that a group satisfying this stronger condition has property A. Since a torsion-free group has the property trivially, we make the following definition. A group has property A if it has elements of finite order and if every proper normal subgroup of finite index is torsion-free.

The question arises: which F-groups have property A?

AMS subject classification, Primary 20H10; Secondary 20E99.

*University of Maryland and National Bureau of Standards.

[1] Figures in brackets indicate the literature references at the end of this paper.

THEOREM: *Let* G *be an* F*-group. Necessary and sufficient that* G *have property* A *is that* G *be a triangle group with distinct prime exponents. Explicitly, this means that in (1.1) we have*

$$g=t=0, \qquad s=3$$

and the $\{l_i\}$ *are distinct primes.*

The presentation of the triangle group can of course be written more compactly as

$$\{e_1, e_2 | e_1^{l_1} = e_2^{l_2} = (e_1 e_2)^{l_3} = 1\}.$$

2. For purposes of this paper we define an F-group as one that has a presentation (1.1) with condition (1.2) satisfied. The following facts are well known. Every F-group G has a faithful representation by a Fuchsian group of the first kind $\bar{G}$:

$$\bar{G} = \Big\{ E_1, \ldots, E_s, P_1, \ldots, P_t, A_1, B_1, \ldots, A_g, B_g | E_1^{l_1} = \ldots = E_s^{l_s} = \prod_i E_i \prod_j P_j \prod_k A_k B_k A_k^{-1} B_k^{-1} = 1 \Big\}, \qquad (2.1)$$

in which E_i is an elliptic, P_j is a parabolic, and A_k, B_k are hyperbolic elements of $SL(2, R)$. Every element in $\bar{G}$ of finite order has order dividing an l_i and is conjugate over $\bar{G}$ to a nontrivial power of E_i. If $\bar{G}$ has an element of finite order, this element is elliptic and has a fixed point lying in the upper half-plane which is a vertex of a fundamental region of $\bar{G}$. This implies that G has an elliptic generator, so that if $\bar{G}$ has an element of finite order, then $s > 0$.

3. LEMMA 1: *If* G *has property* A, *then* $g = t = 0$ *and* $s \geq 3$ *in (2.1)*

Since G has elements of finite order, $s > 0$. Suppose $g > 0$. Let N be the kernel of the homomorphism $G \to Z_2$ sending $a_1 \to \theta$ and all other generators to 1, where θ is a generator of Z_2, the cyclic group of order 2. Then $[G:N] = 2$ and N contains e_1, an element of finite order l_1. Hence G does not have property A.

Therefore $g = 0$. By (1.2) we have

$$s + t \geq 3.$$

Suppose $t > 0$. G is isomorphic to a free product G_1, since p_1 may be eliminated:

$$G \simeq G_1 = \{e_1, \ldots, e_s, p_2, \ldots, p_t | e_1^{l_1} = \ldots = \ldots = e_s^{l_s} = 1\}.$$

If $t \geq 2$, $p_t \notin G_1$. Map $p_t \to \theta$, a generator of Z_2, and all other generators of G_1 to 1; the kernel of this map contains e_1. If $t = 1$, we must have $s \geq 2$; G is then the free product of s cyclic groups of finite order. In that case map $e_1 \to \theta$, a generator of Z_{l_1}, and all other generators of G_1 to 1. The kernel of this map contains e_2.

This completes the proof. G is now of the form

$$G = \}e_1, \ldots, e_s | e_1^{l_1} = \ldots = e_s^{l_s} = e_1 \ldots e_s = 1\}, \; s \geq 3. \qquad (3.1)$$

LEMMA 2: *If* G *has property* A, *the exponents* l_i *in (3.1) are distinct primes, and* $s = 3$.

Suppose, say, $l_1 = mn$, $m > 1$, $n > 1$. Let N be the normal closure in G of e_1^m and let $H = G/N$. A presentation for H is

$$\{e_1, \ldots, e_s | e_1^m = e_2^{l_2} = \ldots = e_s^{l_s} = e_1 \ldots e_s = 1\}.$$

Since $s > 2$, $H \neq \{1\}$. By making use of the theorem of Fox (see [2]), we can deduce the existance of a proper normal subgroup H_1 of H of finite index. If we denote by G_1 the inverse image of H_1 by canonical homomorphism, then G_1 is a proper normal subgroup in G of finite index and G_1 contains e_1^m, an element of finite order n. Hence G does not have property A.

It follows that all l_i are prime. Recall from Lemma 1 that there are at least 3 generators e_i. If $l_1 = l_2$, say, map $e_1 \to \theta$, $e_2 \to \theta^{-1}$, and all other generators to 1, where θ is a generator of $Z_{l_1} = Z_{l_2}$. The kernel of this map contains e_1, $i \neq 1, 2$. Hence the $\{l_i\}$ are distinct.

Finally, suppose $s > 3$. Let N be the normal closure of e_1 in G; let $H = G/N$. H has the presentation $\{e_2, \ldots, e_s \mid e_2^{l_2} = \ldots = e_s^{l_s} = e_2 \ldots e_s = 1\}$. As before $H \neq \{1\}$, the number of independent generators being $s - 2 > 1$, and has elements of finite order. Reasoning as above, we conclude the existence of a proper normal subgroup G_1 of finite index in G and G_1 contains e_1, a contradiction.

Lemmas 1 and 2 complete the necessity part of the Theorem. To prove the sufficiency, consider

$$G = \{x, y \mid x^p = y^q = (xy)^r = 1\}$$

with distinct primes p, q, r. Let N be any normal subgroup of G that contains an element of finite order, u. Then u is conjugate to some nontrivial power of x, y, or xy, say x^α, $0 < \alpha < p$, and since N is normal, x^α itself lies in N. Since p is a prime, $x \in N$. Now G/N is generated by the images $\bar{x}, \bar{y}$ of x, y under the homomorphism $G \to G/N$. Then $\bar{x} = 1$ and so $\bar{y}^q = (\bar{x}\bar{y})^r = \bar{y}^r = 1$, implying $\bar{y} = 1$ since $(q, r) = 1$. Hence $G/N = \{1\}$ or $N = G$. This completes the proof.

References

[1] Bundgard, S., and Nielsen, J., On normal subgroups with finite index in F-groups, Matematisk Tidsskrift **B**, 56–58 (1951).

[2] Fox, R. H., On Fenchel's conjecture about F-groups, Matematisk Tidsskrift **B**, 61–65 (1952).

(Paper 74B4–332)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

Some Theorems on Tensor Composite Graphs*

Michael F. Capobianco**

(September 15, 1970)

If a graph (digraph) is isomorphic to the tensor product of two graphs (digraphs) it is said to be a tensor composite graph (digraph). If not, it is said to be tensor prime. Several theorems giving various properties of tensor composite graphs and digraphs are presented. Among those dealing with (undirected) graphs is the result that any tree is tensor prime. This does not hold for digraphs. An example is given of a tensor composite digraph which is an unoriented tree. It is proved that a tensor composite digraph which is an oriented tree (an arborescence) does not exist. Some applications are presented.

Key words: Digraphs; graphs; products; tensor; trees.

1. Introduction

We refer to Ore [8],[1] Berge [1], and Harary et al. [4], for the usual definitions of graph, digraph (directed graph) and related terms. The tensor product of two graphs G_u and G_v is denoted by $G_u \otimes G_v$ and defined as follows: If V_u and V_v are the sets of vertices of G_u and G_v respectively, then the set of vertices of $G_u \otimes G_v$ is $V_u \times V_v$. Two vertices (u_1, v_1), (u_2, v_2) of $G_u \otimes G_v$ are adjacent if and only if u_1 and u_2 are adjacent in G_u, *and* v_1 and v_2 are adjacent in G_v. If G_u and G_v are digraphs then (u_1, v_1) is adjacent to (u_2, v_2) if and only if u_1 is adjacent to u_2 and v_1 is adjacent to v_2.

A number of papers, [2, 5, 6, 9], have studied various problems of characterizing the component graphs G_u and G_v so that $G_u \otimes G_v$ has certain properties. In this paper, as in an earlier one [3], we focus our attention on the structure of any graph which is isomorphic to the tensor product of two graphs, and investigate its properties. Such a graph (digraph) is called *tensor composite*. If there do not exist graphs G_u and G_v such that $G \cong G_u \otimes G_v$ then G is said to be *tensor prime*.

2. Main Theorems

Our first theorem lists a number of necessary conditions for a graph to be tensor composite. We require the following definition of the distance between two lines of a graph. Let l_1 be the line joining u and v (written $u \sim v$), and let l_2 be $w \sim x$. The *distance* between l_1 and l_2, denoted by $d(l_1, l_2)$, is the length of the shortest path joining any vertex of l_1 with any vertex of l_2.

THEOREM 1: *Let* G *be a tensor composite graph with* n *vertices and* l *lines* ($l > 0$) then (1) $n = n_u n_v$ *where* n_u *and* n_v *are integers greater than* 1 *such that for all vertices* v *of* G, $d(v) \leq (n_u - 1)(n_v - 1)$ *where* $d(v)$ *is the degree of* v, (2) $l \leq 2\binom{n_u}{2}\binom{n_v}{2}$ (3) $l = 2k$, for some $k = 1, 2, \ldots, \binom{n_u}{2}\binom{n_v}{2}$ *and these* l *lines can be listed in* k *pairs such that for each such pair* l_1, l_2, $d(l_1, l_2) > 1$.

PROOF: (1) Let G be isomorphic to $G_u \otimes G_v$ and let n_u and n_v be the number of vertices of G_u and G_v, respectively. Then $n_u > 1$, $n_v > 1$, for otherwise $l = 0$, and furthermore it is clear that $n_u n_v = n$. For the vertices of $G_u \otimes G_v$ we use the abbreviation $(u_i, v_j) = u_i v_j$. Now since $u_i \neq u_i$ and $v_i \neq v_i$ for any i,

---

AMS subject classification. Primary 05C20, Secondary 05C05.

*An invited paper.

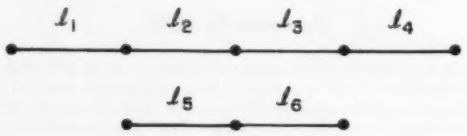
**Present address: St. John's University, Jamaica, New York 11432.

[1] Figures in brackets indicate the literature references at the end of this paper.

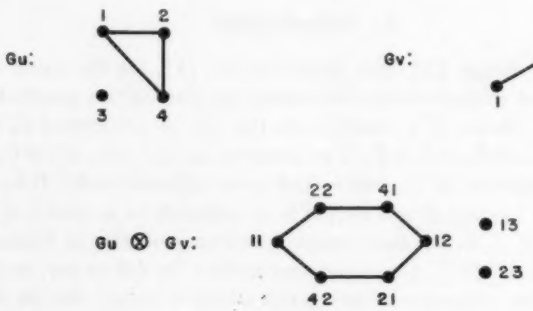
u_iv_j is adjacent with at most $(n_u-1)(n_v-1)$ vertices. Hence $d(v) \leq (n_u-1)(n_v-1)$ for all v in G. (2) $l=\frac{1}{2}\sum_{v\in G} d(v) \leq \frac{1}{2}\sum_{v\in G}(n_u-1)(n_v-1) \leq \frac{1}{2}n_un_v(n_u-1)(n_v-1)=2\binom{n_u}{2}\binom{n_v}{2}$. (3) Suppose $u_1v_1 \sim u_2v_2$. Then $u_1v_2 \sim u_2v_1$. Also $u_1v_1 \neq u_1v_2$, $u_1v_1 \neq u_2v_1$, $u_2v_2 \neq u_1v_2$, $u_2v_2 \neq u_2v_1$. Hence the distance between these two lines is greater than 1.

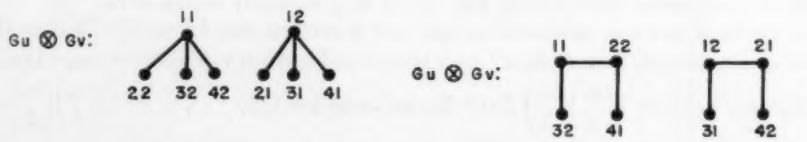
The above theorem has a parallel for digraphs with a similar proof [3].

Note that if l is the number of lines in a tensor composite graph and l_u and l_v are the numbers of lines in the two components graphs then $l=2l_ul_v$. We use this fact in an example which shows that the conditions in theorem 1 are not sufficient, Let G be the graph shown below.



Then $n=8=4\times 2$, $d(v) \leq 2 \leq 3$, $l=6$, $d(l_1, l_4)>1$, $d(l_2, l_6)>1$, $d(l_3, l_5)>1$. However, for G to be tensor composite G_u, say, would have to have 4 vertices and 3 lines, and G_v, 2 vertices and 1 line. The only possibilities are shown below, and G is none of these.





The problem of characterizing tensor composite graphs and digraphs seems quite complex and remains unsolved. A characterization of digraphs having a prime number of lines was given in [3]. The remainder of this paper will be devoted to two nonexistence theorems and their applications.

THEOREM 2: *No tree is tensor composite.*

PROOF: We use two lemmas.

LEMMA 1: *If* G_u *or* G_v *has a cycle, then so does* $G_u \otimes G_v$.

PROOF: Let

$$u_1 \sim u_2 \sim u_3 \sim \ . \ . \ ., \sim u_k \sim u_1, v_1 \sim v_2.$$

Then if k is even,

$$u_1v_1 \sim u_2v_2 \sim u_3v_1 \sim u_4v_2 \sim v_3v_1 \sim \ . \ . \ . \sim u_kv_2 \sim u_1v_1$$

so that $G_u \otimes G_v$ has a cycle. If k is odd then

$$u_1v_1 \sim u_2v_2 \sim u_3v_1 \sim u_4v_2 \sim \ . \ . \ . \sim u_kv_1 \sim u_1v_2 \sim u_2v_1 \sim \ . \ . \ . \sim u_kv_2 \sim u_1v_1$$

so that $G_u \otimes G_v$ again has a cycle.

LEMMA 2: *The tensor product of two forests is not connected.*

PROOF: This lemma follows, in part, from Weichel's theorem 1, [9], which states that a tensor composite graph is connected if and only if the component graphs are connected and at least one of them has an odd cycle. For the sake of completeness, we give an independent proof of the lemma here.

Consider u_1v_1 and u_1v_2 with $v_1 \sim v_2$. We claim that there is no path between u_1v_1 and u_1v_2. There can not be a path between u_1 and u_1 since this is a cycle and its existence would contradict the fact that G_u is a forest. Therefore, the only way one can have a path between u_1v_1 and u_1v_2 is if this path has the form

$$u_1v_1 \sim u_2v_j \sim u_1v_k \sim u_2v_m \sim \ . \ . \ . \ . \sim u_1v_2.$$

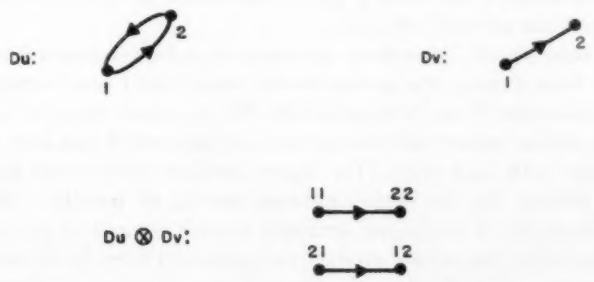
We examine the possibilities. Of course $j \neq 1$, $k \neq 1$. If $j=2$ then $k \neq 2$. Let $k=3$. Then $m \neq 1$ for otherwise there would exist a cycle containing v_1. It is also clear that $m \neq 2$. Therefore, $m=4$. But this implies the existence of a cycle containing v_2. Hence we must have $j=3$. However, this implies the existence of a path between v_1 and v_2 other than $v_1 \sim v_2$, which would mean that v_1 and v_2 are contained in a cycle. This final contradiction proves lemma 2.

Since a tree is a connected graph without cycles, the theorem follows immediately.

Lemma 2 has a parallel for digraphs. One talks about semicycles instead of cycles, and weakly connected instead of connected. Thus if we understand that a forest is a digraph without semicycles than we can prove the following result.

LEMMA 2′: *The tensor product of two forests is not weakly connected.*

Lemma 1, however does not hold for digraphs, as can be seen by the following example.



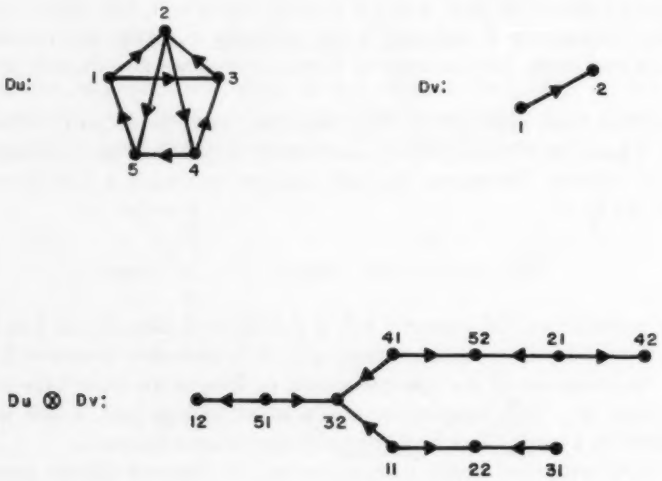
Hence the proof of theorem 2 does not go through for digraphs. In fact, the theorem is not true, as can be seen by the example below.

Note that in this example the tree obtained is an unoriented one. One immediately asks if an oriented tree (an arborescence) can be tensor composite. The answer is given in the next theorem.

THEOREM 3: *Except for the trivial graph (a single point), any arborescence is tensor prime.*

PROOF: Consider a nontrivial tensor composite digraph D in which u_1v_1 has indegree 0. Then $u_iv_j \neq u_1v_1$ for all i and j. Therefore, either $u_i \neq u_1$ for all i or $v_j \neq v_1$ for all j. Hence for some $k \neq 1$ either u_1v_k or u_kv_1 has indegree 0. Since an arborescence requires exactly one point with *indegree* 0, D can not be an arborescence.

Using a dual argument it is easy to see that no nontrivial tensor composite digraph can have exactly one point with *outdegree* 0.



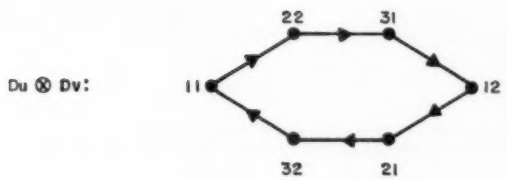
3. Some Applications

We present applications of theorems 2 and 3. Theorem 2 is used in the form: Any connected tensor composite graph has at least one cycle.

Consider a message which for security purposes is to be transmitted in two parts, possibly a text and a key. We have a group of n agents among whom part 1 can be transmitted, and a group of m agents among whom part 2 can be transmitted. We represent these two groups by two graphs in which the vertices are the agents and two vertices are adjacent if and only if the two agents they represent communicate with each other. The tensor product of these two graphs then represents the communication pattern for the entire message among all possible "teams" of two agents each. Theorem 2 tells us that if things are arranged in such a way that the message can get from any agent team to any other (the tensor product is connected) then there must exist at least three teams which can experience feedback in the sense that a message can be returned to them (the tensor product has at least one cycle).

Theorem 3 can be applied to a situation in which two sets of players compete in tournaments. We represent each set by a digraph in which the vertices are the players and u is adjacent to v if and only if u defeats v. In the case of a tie we let u and v be adjacent to each other. Now form all possible teams, or coalitions, of two players, one from the first set and one from the second set. Assuming that the strengths of the players are additive, the tensor product will represent the "victory-defeat" pattern for these coalitions. From the proof of theorem 3 we see that there can

not be exactly one undefeated coalition. Either there is none, or there are more than one. That the latter is possible can be seen by our last two examples. The former is illustrated below.



4. References

[1] Berge, C., Theory of Graphs and its Applications (Wiley 1962).

[2] Capobianco, M. F., Tournaments and tensor products of digraphs, SIAM J. App. Math. **15**, 624–626 (1967).

[3] Capobianco, M. F., Tensor products of digraphs and the structure of groups of pairs, Bull. Math. Biophysics **31**, 319–326 (1969).

[4] Harary, F., Norman, R., Cartwright, D., Structural Models (Wiley 1965).

[5] Harary, F., and Trauth, C., Connectedness of products of two directed graphs, SIAM J. App. Math. **14**, 250–254 (1966).

[6] McAndrew, M., On the product of directed graphs. Proc. Amer. Math. Soc. **14**, 600–606 (1963).

[7] Miller, D., The categorical product of graphs, Can. J. Math. **20**, 1511–1521 (1968).

[8] Ore, O., Theory of graphs, Amer. Math. Soc. (1962).

[9] Weichsel, P., The kronecker product of graphs, Proc. Amer. Math. Soc. **13**, 47–52 (1962).

(Paper 74B4–333)

JOURNAL OF RESEARCH of the National Bureau of Standards – B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

The Diophantine Approximation of Roots of Positive Integers*

Charles F. Osgood**

(September 21, 1970)

The following result is established:
THEOREM: *Suppose that* $k \geq 150$ *and* m *are fixed positive integers. Then*

$$|\sqrt[k]{m} - pq^{-1}| < q^{-\frac{7}{8}k}$$

can hold for at most one pair of relatively prime positive integers p *and* q *with* $q \geq 2^9(\sqrt[k]{m}+1)^6$.
The new feature of this result is that the lower bound on q is given explicitly and is "small."

Keywords: Diophantine approximation; diophantine equation; effective computability.

1. Introduction

Recently Schinzel [4][1] and Davenport [1] have each obtained a result of the following sort: Let α be a real algebraic number of degree $r \geq 2$. Let s be a positive real number larger than $s(r)$, where for Davenport $r > s(r) = \frac{1}{2}r + 0(1) > \frac{1}{2}r$ while for Schinzel $s(r) = 3(r/2)^{1/2}$. Then there exists an effectively computable positive integer $q_0(\alpha, s)$ such that, with at most one exception, every pair of relatively prime integers p and q with $q \geq q_0(\alpha)$ satisfies the inequality

$$|\alpha - pq^{-1}| \geq \tfrac{1}{2}q^{-s}.$$

(Also Seppo Hyyrö in [2] obtained something analogous to Davenport's result for kth roots of rational numbers, where $k \geq 2$ is a positive integer.)

None of these authors, however, calculated $q_0(\alpha, s)$ explicitly for any class of α and s. With the aid of a theorem in a recent paper by the present author we can obtain explicitly a rather "small" $q_0(\alpha, s)$ for a certain class of α and s. (Below one could drop the lower bound on k very considerably by allowing a larger $s < k$ and a larger $q_0(\alpha, s)$.)

THEOREM I: *Suppose that* $k \geq 150$ *and* m *are fixed positive integers. Then*

$$|\sqrt[k]{m} - pq^{-1}| < q^{-\frac{7}{8}k} \tag{1}$$

can hold for at most one pair of relatively prime positive integers p *and* q *with* $q \geq 2^9(\sqrt[k]{m}+1)^6$.

DEFINITIONS: *By a reduced approximation we shall mean a fraction* $\frac{p}{q}$ *where* p *and* q *are positive integers and* $(p, q) = 1$. *Set* $\theta(\beta) = \frac{k}{6}\left((\frac{7}{8}k - 2)^{\beta} - 15\right)$.

AMS subject classification. Primary 10F05.

*An invited paper. This paper was written in part while the author was on a postdoctoral Research Associateship at the National Bureau of Standards, Washington, D.C. (This award is given by the National Bureau of Standards in association with the National Academy of Sciences and the National Research Council.)

**Present address: Naval Research Laboratory, Washington, D.C. 20390.

[1] Figures in brackets indicate the literature references at the end of this paper.

THEOREM II: *If* m *and* $k \geq 150$ *are positive integers and* β *is any positive integer such that* $m < 2^{6(\beta)}$ *then* (1) *is satisfied by at most* $\beta+1$ *distinct reduced approximations.*

THEOREM III: *If* $k \geq 150$ *and* $m < (q_0^{\frac{7}{8}k-1}2^{-16})^{\frac{k}{6}}$ *are positive integers then* (1) *is satisfied by at most two distinct reduced fractions with denominators larger than or equal to* $q_0 \geq 2$.

Setting $q_0 \geq 10$ in Theorem III we obtain,

COROLLARY: *If* $k \geq 150$ *and* m *are positive integers with* $\sqrt[k]{m} < 10^{20}$ *then* (1) *is satisfied by at most two distinct reduced fractions with denominators larger than 9.*

2. Section I

PROOF OF THEOREM I. *We shall use the Theorem of* [3] *(which will be stated below for the case* $k_1=1$, $n=2$, $\epsilon=2$*) and the supposed existence of two solutions of* (1), p', q' *and* p'', q'' *where* $(p', q')=(p'', q'')=1$ *and* $q'' > q' \geq 2^9(\sqrt[k]{m}+1)^6$.

From [3] we have: Let s and k be positive integers, and $k \geq 2$. Let $0 < \epsilon < +\infty$ be a real number. Let

$$K = 2^{\frac{3}{2}k+\frac{1}{2}}s^2.$$

Let N denote a positive integer larger than K. Set

$$1 > \Lambda(N) = \frac{\log(NK^{-1})}{\log(2^7NK)} > 0,$$

and

$$\varphi(N) = (2^7KN)^{6+3k^{-1}}.$$

Let q denote a positive integer and (p_1, p_2) a nonzero vector of nonnegative integers. Let C denote any real number satisfying $0 \leq C \leq 1$. Then we have,

THEOREM: If $q > \varphi(N)$

$$\max\{|CN^{k-1} - p_1q^{-1}|, |C(N+s)^{k-1} - p_2q^{-1}|\} \geq \tfrac{1}{2}(2q)^{-\left(1+\frac{3}{\Lambda(N)}\right)}$$

for all C and (p_1, p_2).

Returning to the proof of Theorem I, if $|\sqrt[k]{m} - p'(q')^{-1}| < (q')^{-\frac{7}{8}k}$ then $|(q')^k m - (p')^k| < k(\sqrt[k]{m}+1)^{k-1}(q')^{\frac{k}{8}}$. Choose N to be the smaller of $m(q')^k$ and $(p')^k$. (Then $m(q')^k \geq N > m(q')^k - k(\sqrt[k]{m}+1)^{k+1}(q')^{\frac{k}{8}}$.) Set $s = |(q')^k m - (p')^k|$.

Note that then

$$K < 2^{\frac{3}{2}k+\frac{1}{2}}k^2(\sqrt[k]{m}+1)^{2k-2}(q')^{\frac{k}{4}}. \tag{2}$$

We shall presently show that $N > K$. Assuming $N > K$ for the present and setting $C=1$, we see that by the Theorem from [3],

$$|\sqrt[k]{m} - pq^{-1}| \geq (2q')^{-1}(2q)^{-\left(1+\frac{3}{\Lambda(N)}\right)} \tag{3}$$

for all positive integers p and q with $q > (2^7KN)^{6+3k^{-1}}$. We shall eventually contradict (3) with $p = rp''$, $q = rq''$ for a positive integer $r \geq 2$.

First to obtain lower bounds on N. Using $q' > 2^9(\sqrt[k]{m}+1)^6$ and $k \geq 150$ we see that

$$m(q')^k > 2(2^{3k+1}k^4(\sqrt[k]{m}+1)^{4k-4})(q')^{k/2}.$$

Also, trivially,

$$k(\sqrt[k]{m}+1)^{k-1}(q')^{k/8} < 2^{3k+1}k^4(\sqrt[k]{m}+1)^{4k-4}(q')^{k/2}.$$

Thus

$$N > m(q')^k - k(\sqrt[k]{m}+1)^{k-1}(q')^{k/8} > 2^{3k+1}k^4(\sqrt[k]{m}+1)^{4k-4}(q')^{k/2} > K^2 > K. \tag{4}$$

The last two inequalities come from (2).

Using the triangle inequality we conclude that

$$2(q')^{-\frac{7}{8}k} > \left|\frac{p'}{q'} - \frac{p''}{q''}\right| > (q'q'')^{-1}; \qquad \text{thus,} \qquad q'' > \tfrac{1}{2}(q')^{\frac{7}{8}k-1} \overset{\text{def.}}{=} M_1.$$

Now

$$1 + \frac{3}{\Lambda(N)} = 4 + \frac{3\log(2^7K^2)}{\log(NK^{-1})} < 4 + \frac{3\log(2^7K^2)}{\log K}$$

because $N > K^2$ as we saw in (4). Thus

$$1 + \frac{3}{\Lambda(N)} < 10 + \frac{3\log 2^7}{\log K} < 10\frac{1}{10}, \tag{5}$$

since $K = 2^{\frac{3}{2}k+\frac{1}{2}}s^2 > 2^{225}$. Now since $q' < q''$ and $k \geq 150$,

$$(2q')^{-1}(2q'')^{-10.1} > (q'')^{-\frac{7}{8}k}$$

so by the Theorem from [3], by formula (2), and by our bound on N,

$$q'' < (2^7KN)^{6+3k^{-1}} \leq (m(\sqrt[k]{m}+1)^{2k-2}2^{7\frac{1}{2}+\frac{3}{2}k}k^2(q')^{\frac{5}{4}k})^{6+3k^{-1}} \overset{\text{def.}}{=} M_2.$$

Let [2] $r = [M_2M_1^{-1}] + 1$ and $\delta = \dfrac{\log(q'')}{\log(2rq'')}$. Now $\left|\sqrt[k]{m} - \dfrac{rp''}{rq''}\right| < (2rq'')^{-\frac{7}{8}\delta k}$.

Further $\delta \geq [\log(q'')][\log(4M_2M_1^{-1}q'')]^{-1} > [\log(M_1)][\log(4M_2)]^{-1}$, since $q'' > M_1$.

We wish to show that

$$[\tfrac{7}{8}k\log(\tfrac{1}{2}(q')^{\frac{7}{8}k-1})][\log(4m(\sqrt[k]{m}+1)^{2k-2}2^{7\frac{1}{2}+\frac{3}{2}k}k^2(q')^{\frac{5}{4}k})^{6+3k^{-1}})]^{-1} \geq 10\tfrac{1}{8}. \tag{6}$$

This would give us our desired contradiction since by (3) and (5) and the inequalities $q'' > \frac{1}{2}(q')^{\frac{7}{8}k-1}$ and $k \geq 150$

$$\left|\sqrt[k]{m} - \frac{rp''}{rq''}\right| > (2q')^{-1}(2rq'')^{-10.1} > (2rq'')^{-10.2}$$

while we would have that

$$\left|\sqrt[k]{m} - \frac{rp''}{rq''}\right| < (2rq'')^{-10.2}.$$

Since $m(\sqrt[k]{m}+1)^{2k-2} < (q')^{\frac{k}{2}}2^{-4\frac{1}{2}k}$ and $k^2 < 2^k$ we see that (6) is implied by

$$\tfrac{7}{8}k[\log(\tfrac{1}{2}(q')^{\frac{7}{8}k-1})][\log 4((q')^{\frac{7}{4}k}2^{-2k+7\frac{1}{2}})^{6+3k^{-1}})]^{-1} \geq 10\tfrac{1}{8},$$

[2] Here only we use the greatest integer notation.

which is implied by

$$[\tfrac{7}{8}k \log ((\tfrac{1}{2}q')^{\frac{7}{8}k-1})][\log ((\tfrac{1}{2}q')^{\frac{7}{4}k(6+3k^{-1})}]^{-1} \geq 10\tfrac{1}{8}.$$

But then all that we have to see is that

$$\tfrac{1}{2}(\tfrac{7}{8}k-1)(6+3k^{-1})^{-1} \geq 10\tfrac{1}{8}$$

if $k \geq 150$. This is easily done and proves Theorem I.

3. Section II

PROOF OF THEOREM II: *If* $\frac{p'}{q'}$ *and* $\frac{p''}{q''}$ *satisfy* (*1*) *and* $q'' > q'$ *then either* $\frac{p'}{q'} = \frac{p''}{q''}$ *or* $q'' > \frac{1}{2}(q')^{\frac{7}{8}k-1}$. *Thus if there is a* $\beta+1-$st *reduced approximation* (*ordering them by the magnitude of their denominators*) *it has a denominator at least as large as* $2^{(\frac{7}{8}k-2)^{\beta}} \geq 2^{15}(\sqrt[k]{m})^6 \geq 2^9(\sqrt[k]{m}+1)^6$, *so it is the final reduced approximation which satisfies* (*1*).

PROOF OF THEOREM III: If there are two reduced fractions $\frac{p'}{q'}$ and $\frac{p''}{q''}$ with $q'' > q' \geq q_0$ which satisfy (1) then

$$q'' > \tfrac{1}{2}q_0^{\frac{7}{8}k-1} > 2^{15}(\sqrt[k]{m})^6 \geq 2^9(\sqrt[k]{m}+1)^6,$$

so these are all of the reduced approximations which satisfy (1).

4. References

[1] Davenport, H., A Note on Thue's Theorem, Mathematika, vol. **15**, 76–87 (1968).
[2] Hyyrö, Seppo, Über Die Catalansche Problem, Ann. Acad. Scient. Fennicae, Series **A**, No. 355, 1–50 (1964).
[3] Osgood, C. F., The Simultaneous Approximation of Certain k-th Roots, Proc. Camb. Phil. Soc. **67**, 75–86 (1970).
[4] Schinzel, Zentralblatt für Math., **137**, 258 (1967).

(Paper 74B4–334)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Science
Vol. 74B, No. 4, October–December 1970

The Minimum Number of Problems to Cover All Subproblems*

H. J. Greenberg**

(June 24, 1970)

The following problem is motivated and then solved, using the theory of systems of distinct representatives. Let $M=\{1, 2, \ldots, m\}$, and for each sequence σ of distinct members of M, let $\langle\sigma\rangle$ be the associated subset. Suppose given a mathematical problem $P(S)$ for each subset S of M, and an algorithm A which when applied to σ solves not only $P(\langle\sigma\rangle)$ but also all $P(\langle\tau\rangle)$ where τ is an initial segment of σ. What is the smallest number of applications of A needed to solve the entire ensemble of problems $\{P(S):S \subset M\}$?

Key words: Combinatorics; dynamic programming; optimization; systems of distinct representatives; theory of computation.

1. Introduction

This note is presented with the same intent as a note by A. J. Goldman [4],[1] where the motive is "to illustrate how mathematical disciplines that appear too abstruse to be useful to the operations researcher can in fact provide exactly what is needed to answer some of his questions." The particular mathematical discipline used in our analysis is the combinatorial theory of "distinct representatives."

Let $M=\{1, 2, \ldots, m\}$, and suppose a mathematical problem $P(S)$ is associated with each subset S of M. We shall be concerned with a measure of the computational labor required to solve the ensemble of problems $\{P(S):S \subset M\}$, under the especially favorable circumstances described below. Obtaining the ensemble of solutions will be referred to as *covering all subproblems.*

For any sequence σ of distinct members of M, let $\langle\sigma\rangle$ denote the subset of M whose members are the terms of σ. Suppose we have an algorithm A, which when applied to σ solves not only the problem $P(\langle\sigma\rangle)$ but also all the problems $P(\langle\tau\rangle)$ where τ is an initial segment of σ. Our question is: What is the minimum number $N(m)$ of applications of A needed to cover all subproblems? To which sequences should A be applied to achieve this minimum number?

The analysis of this question is deferred to the next section of the note. We first present several examples to show how this question arises in mathematical operations-research settings.

Example 1. Consider a 0–1 integer linear program with m binary variables, and with objective function c^Tx. There are 2^m binary m-vectors to consider. Methods of implicit enumeration [1, 3] evaluate the objective function for some subfamily of these 2^m arguments, and the efficiency of such a method is typically measured by comparing the size of this subfamily with the number 2^m of evaluations required in a brute-force approach.

We propose, however, that an appropriate basis for comparison may be not 2^m but rather the smaller value $N(m)$ defined above. To see why, observe that the binary m-vectors x are in natural correspondence with the subsets S of M; let $x(S)$ be the vector associated with S, and let $P(S)$ be the task of evaluating $c^Tx(S)$. This evaluation involves, for some arrangement of the members of

AMS subject classification: Primary 05A05.
*An invited paper.
**Present address: Computer Sciences Center, Southern Methodist University, Dallas, Texas 75222.
[1] Figures in brackets indicate the literature references at the end of this paper.

S into a sequence σ, the formation of the successive partial sums $c_{\sigma(1)}$, $c_{\sigma(1)}+c_{\sigma(2)}$, etc. Thus we have an instance of the general question under study. The same idea applies when $P(S)$ is the evaluation of some function $F(S)$ which would naturally be calculated using a recursion

$$F(T \cup \{j\}) = F(T) * f(j) \qquad (j \in M - T)$$

where "*" is a commutative operation and f is a function defined on M.

Example 2. As another example, suppose we have an m-stage dynamic program, the stages being numbered by the members of M. Assume this program has the special property that its solution is independent of the ordering of stages, and that the same is true of each of the subproblems $P(S)$ defined by using only some subset S of the stages. For certain sensitivity analyses, it is desirable to solve the entire ensemble $\{P(S): S \subset M\}$ of problems; the recursive nature of the standard dynamic algorithm makes this an instance of the general question. For example, if $m = 3$, it will follow from our analysis that the minimum number of dynamic programs which must be solved is 3. One such trio is:

Program 1: Stage 1 → Stage 2 → Stage 3
Program 2: Stage 2 → Stage 3
Program 3: Stage 3 → Stage 1

Applying the dynamic programming algorithm to Program 1 yields the solutions of $P(\{1\})$, $P(\{1, 2\})$, and $P(\{1, 2, 3\})$; application to Program 2 solves $P(\{2\})$ and $P(\{2, 3\})$; and application to Program 3 yields the remaining nontrivial members $P(\{3\})$ and $P(\{1, 3\})$ of the problem ensemble.

An example of a dynamic programming model having the stated commutivity property on its its stages is an additive resource allocation problem of the form:

$$\text{Max} \sum_{j=1}^{m} f_j(x_j) : \sum_{j=1}^{m} g_j(x_j) \leq b$$

$$x_j \in \{0, 1, \ldots, d_j\}, \qquad j = 1, \ldots, n.$$

The problem, $P(S)$, is of the form

$$\text{Max} \sum_{j \in S} f_j(x_j) : \sum_{j \in S} g_j(x_j) \leq b$$

$$x_j \in \{0, 1, \ldots, d_j\}, \qquad j \in S.$$

Example 3. The writer first encountered the general problem in the analysis of a parametric min-cut problem for communications networks [5]. In that context we are given a finite network with a set N of *nodes* and a set A of directed *arcs*; there is a distinguished *source* node s, and a distinguished subset T of $N - \{s\}$ consisting of m *terminal* nodes. A *cut* between s and a subset, T', of T is a subset of A whose removal from the network would leave intact no path from the source s to any node of T'.

In this application we are also given a nonnegative rational-valued *cost* function c defined on the set A of arcs, and a nonnegative rational-valued *value* function v defined on the set T of terminals; c is extended additively to the subsets of A (e.g., cuts), and v is extended additively to the subsets T' of T. The parametric analysis requires, for each number V in the interval $[0, v(T)]$, determining among those subsets T' of T with $v(T') \geq V$, one which minimizes $f(T') = \min [c(C)]$ where C ranges over all cuts between s and T'.

Such an analysis can be carried out in many ways [5]. The one which concerns us here involves the determination, for each of the 2^m subsets T' of T, of the associated value $f(T')$ (as well as the relatively trivial calculation of $v(T')$). Evaluation of a given $f(T')$ can be accomplished by

applying the Ford-Fulkerson labelling algorithm [2]. This algorithm, however, has an inherent ability to solve subproblems in a recursive manner. That is, the "labels" determined in evaluating $f(T')$ can be used in solving for $f(T' \cup \{t\})$ where $t \epsilon T - T'$. Thus, to determine the minimum number of explicit applications of the min-cut algorithm needed to complete the analysis while exploiting this recursive property, is again an instance of our general question.

2. Analysis

It is convenient to translate the general question into the setting of an abstract network, whose nodes are the 2^m subsets of M, and whose arcs are the pairs (S, S') such that $S' = S \cup \{j\}$ for some $j \epsilon M - S$. In this setting, an arrangement of the members of S into a sequence σ is equivalent to selecting a specific path from node ϕ (corresponding to the null set) to node S; initial segments of σ correspond to initial subpaths of this path. Thus our general question takes the following form:

What is the minimum number, $N(m)$, of paths starting at ϕ, which will cover all the nodes of this abstract network?

Let $[x]$ denote the greatest integer not exceeding x. It will be useful to have a special notation for the largest binomial coefficient with "upper" argument m; let

$$g(m) = \binom{m}{[m/2]}.$$

Our analysis will yield a constructive proof that

$$N(m) = g(m). \tag{2.1}$$

There are $g(m)$ nodes corresponding to $[m/2]$-tuples, and no two such nodes can be on a common path starting from ϕ (i.e., no $[m/2]$-tuple is contained in another). This shows that $N(m) \geq g(m)$. It only remains to show that there is a set of $g(m)$ paths starting at ϕ which together cover all nodes.

The abstract network has an obvious "mirror symmetry" associated with the one-to-one correspondence between sets S and their complements $M - S$. Exploiting this, we see that it suffices to confine attention to the subnetwork containing the nodes corresponding to sets of cardinality $\leq [m/2]$, and to show that all nodes of this subnetwork can be covered by a suitable set of $g(m)$ paths starting at ϕ.

Let "level p" refer to those nodes which correspond to subsets with p members. The construction of the desired $g(m)$ paths can be effected recursively, proceeding from level to level in the network. It suffices to prove, by induction on p for $p \leq [m/2]$, that there are paths from ϕ to the $\binom{m}{p}$ nodes at level p, one path per node at level p, which are *special* in the sense that these paths together cover all nodes at levels $\leq p$. What is required for the induction step is to show, for $p < m/2$, that one element can be deleted from each $(p+1)$-tuple in such a way that the resultant set of p-tuples includes *all* p-tuples. That is, there are one-arc extensions of the special paths for level p (with each of these paths receiving at least one such extension) which together cover all nodes at level $p+1$; these extended paths are then clearly a special family for level $p+1$.

This is established in the following theorem and subsequent corollary, thus completing the proof of (2.1). The theorem will be demonstrated using the theory of distinct representatives, in particular the basic theorem of Hall [7]; an alternative would be the approach adopted by Meyers [8] in treating a combinatorial question somewhat related to that of the present note.

406-680 O - 71 - 2

THEOREM: *Let* $\{T_i\}_1^n$ *be a collection of sets which collectively contain* q *elements, say* $\{e_j\}_1^q$. *Let* T_i *contain* t_i *elements, and assume* e_j *occurs in* r_j *of the sets. Define*

$$\rho = \min_j \{r_j\}_1^q$$

$$\tau = \max_i \{t_i\}_1^n$$

If $q\rho \geq n\tau$ *and* $\rho \leq \tau$, *then the family* $\{T_i\}_1^n$ *has a SDR (system of distinct representatives).*

PROOF: Let I be any index set from $\{1, \ldots, n\}$, with k members. Suppose $\bigcup_{i \in I} T_i$ contains a elements, say $\{e_j\}_{j \in J}$ (J is an index set from $\{1, \ldots, q\}$ with a members). Then, for each $j \in J^c$ (complement of J) none of the r_j repetitions of e_j are in any of $\{T_i : i \in I\}$. Therefore, the number of appearances of elements in $\{T_i : i \in I^c\}$ is at least $\sum_{j \in J^c} r_j$. This yields the inequality:

$$\sum_{j \in J^c} r_j \leq \sum_{i \in I^c} t_i.$$

Therefore,

$$(q-a)\rho \leq (n-k)\tau.$$

Since $q\rho \geq n\tau$ and $\rho \leq \tau$, it follows that $a \geq k$. Our conclusion then follows from Hall's Theorem.

Observe that the counting process reveals $\sum_{j \in J^c} r_j = \sum_{i \in I^c} t_i$. Therefore, if all r_j are equal and all t_i are equal (so $r_j = \rho$ and $t_i = \tau$), and if $\rho \leq \tau$, then the theorem must apply with $q\rho = n\tau$. This says that in the corollary below we rely upon the fact that

$$\binom{m}{p+1}(p+1) = \binom{m}{p}(m-p)$$

instead of merely

$$\binom{m}{p+1}(p+1) \geq \binom{m}{p}(m-p).$$

COROLLARY: *For the set* $M = \{1, \ldots, m\}$, *let* $\{P_i\}_1^k$ *be the collection of all* p*-tuples where* $k = \binom{m}{p}$ *and* $p < m/2$. *Then there exists a sequence of elements* $\{s_i\}_1^k$, *with repetitions allowed, such that* $\{P_i \cup s_i\}_1^k$ *are distinct* $(p+1)$*-tuples.*

PROOF: Let T_i be the collection of $m-p$ $(p+1)$-tuples formed by adjoining to P_i any member of its complement, $M - P_i$. Regard P_i as "representable" by any $(p+1)$-tuple in T_i; then what is to be shown is that $\{T_i\}_1^k$ admits a system of distinct representatives.

In the notation of our theorem we have

$$q = \binom{m}{p+1}, \qquad n = \binom{m}{p},$$

$$r_j = p+1 \text{ for all } j = 1, \ldots, q,$$

$$t_i = m-p \text{ for all } i = 1, \ldots, n.$$

Therefore, $\rho = p+1$ and $\tau = m-p$. Observe that

$$q\rho = n\tau \text{ and } \rho \leq \tau \text{ for } p < m/2.$$

Hence, our proof is complete.

The computation of the $g(m)$ paths is performed level by level by solving for the completions (or deletions) indicated by the corollary. It suffices to consider $p < m/2$ and describe a method to proceed from level p to level $p+1$. In terms of our abstract network our problem is to fined a maxi-

mal matching in the bipartite network consisting of the $\binom{m}{p}$ vertices of level p, the $\binom{m}{p+1}$ vertices of level $p+1$ and the edges that connect a node in level p with a node in level $p+1$.

Equivalently, we can cast our problem as an assignment problem with a cost matrix, C, having only $0-1$ elements. We set $c_{ij}=0$ if node i (in level p) is connected with node j (in level $p+1$); otherwise $c_{ij}=1$. We then ask for an assignment of each row to a distinct column to minimize the sum total cost. We know there is a solution with zero cost, and the assignment provides the desired matching.

The author wishes to thank A. J. Goldman for his many helpful comments.

3. References

[1] Balas, E., An additive algorithm for solving linear programs with zero-one variables, Oper. Res. **13,** 517–546 (1965).
[2] Ford, L. R., Jr., and Fulkerson, D. R., Flows in Networks (Princeton Univ. Press, Princeton, N.J., 1962).
[3] Glover, F., A multiphase-dual algorithm for the zero-one integer programming problem, Oper. Res. **13,** 879–919 (1965).
[4] Goldman, A. J., Fractional container-loads and topological groups, J. Oper. Res. **16,** 1218–1221 (1968).
[5] Greenberg, H. J., Optimal attack of a communications network, Ph. D. Thesis, The Johns Hopkins University (Baltimore, Md., 1968).
[6] Greenberg, H. J., The use of branching in dynamic programming for parametric analysis, Oper. Res. **15,** 976–977 (1967).
[7] Hall, P., On representatives of subsets, J. London Math. Soc. **10,** 26–30 (1935).
[8] Meyers, P. R., Minimum number of subsets to distinguish individual elements, J. Res. Nat. Bur. Stand. (U.S.), **71B** (Math. and Math. Phys.), 21–22 (1967).

(Paper No. 74B4–335)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

Error Estimates for the Solution of Linear Algebraic Systems*

Brother Kenneth E. Fitzgerald, F.S.C.**

(July 20, 1970)

In this paper bounds for the error of a computed inverse of a matrix are developed. These are then applied to the solution of a single system. Methods for improving the approximate inverse are then discussed with some observations on the dangers involved in their practical use on a computer and some safeguards are indicated. Some computer programs for matrix inversion are then evaluated by means of the bounds developed.

Key words: Error estimates; evaluation of computer programs; inverse of a matrix; linear systems and matrices.

Contents

	Page
Introduction	251
1. Error bounds for the approximate inverse. Influence of $N(AB) \leq N(A)N(B)$; Absolute error bounds; Bounds when data is not exact; Relative error bounds; Numerical example.	253
2. Error bounds for the approximate solution vector of $Ax=b$. Absolute error bounds; Remark on the use of the approximate inverse; Relative error bounds; Bounds when data is not exact; Remarks about the need for approximate inverse.	263
3. Improvement schemes for the approximate inverse. Three schemes, absolute error bounds; Generalization, absolute error bounds; Relative error bounds; Comparison; Two standard iterative procedures; Numerical example.	266
4. Improvement schemes for the approximate solution vector $Ax=b$. Standard scheme; Different approach, convergence, error bounds; Different convergence proof for the standard scheme, error bounds.	276
5. Error bounds for the approximate inverse of special matrices. Symmetric; triangular; LU decomposition; Application to solution of $Ax=b$.	282
6. Evaluation of computer programs. Test matrices; Computer programs; Results; Evaluation and discussion.	292
7. References	309
Selected bibliography	309

Introduction

The advent of modern high-speed electronic computers has led to the solution of many problems that had been considered almost unsolvable some years ago. The essential contribution of these computers is that they can do simple things extremely fast. Mathematical problems, like the solution of a system of many linear algebraic equations, can be done in a few seconds whereas it would take many hours by a person with a desk calculator and even longer if he used only pencil and paper. If we actually solved the problem in either of the latter ways, we would have a good idea of the errors we allowed to enter the process and could keep track of them. It is in this area that high-speed computers have some limitations. We feed numbers into a computer and the com-

AMS subject classification. Primary 65F35, Secondary 68A20.

*An invited paper. A dissertation submitted to the faculty of the Graduate School of Arts and Sciences of the Catholic University of America in partial fulfillment of the requirements for the degree Doctor of Philosophy.

**Present address: Manhattan College, Bronx, N.Y. 10471

puter prints out other numbers—supposedly the solution of our problem. But how do we know that these numbers *are* the solution of our problem? That this is a problem is pointed out very well by Leslie Fox:

> In the desk-machine era of Comrie and Peters, the great compilers of mathematical tables, the possibility of errors and the perpetual need for guarding against them were fully realised. They would be appalled by the self-confidence of their successors! [1, p. 16][1]

It certainly is possible that the machine could malfunction; some physical part could be defective—this has, in fact, happened and with no indication to the operator. Some modern text books on numerical analysis do not seem to recognize this possibility and this is most unfortunate [2, p. 111]. We will indicate in section 6 one way of checking for this kind of error.

Another possibility that exists is that our instructions to the machine (the "program") may not accomplish exactly what we had intended. There are many reasons for this, such as human error in punching the deck, the rounding process that continually occurs, subtraction of numbers that are very close to each other, and even poor programming. These difficulties are being investigated very seriously now by many men in the field, especially some in the Special Interest Group in Numerical Mathematics of the Association for Computing Machinery. We must be very sure that the computer program we are using does, in fact, perform successfully. Evaluation of existing computer programs is going on [3] and, in fact, we have devoted the second part of our research to this very thing, but, of course, only in a specific area. The criteria for this evaluation will be discussed at the beginning of section 6 and the search for these criteria gave the impetus for the first part of this research.

In this first part we derive some theoretical bounds for the accuracy of a computed solution of a system of linear algebraic equations. We concentrate first on the problem of finding an error bound for the computed inverse of a matrix (the solution of a set of linear algebraic systems) and then consider the solution of a single system. (This order may seem backwards at first, but we indicate its necessity at the end of section 2.) Our orientation will be to use what is computationally available. That is to say, if we are interested in the accuracy of a computed inverse, X, we will assume that the residual, $Y=I-AX$, is also available. We will then discuss ways of improving the computed inverse and the calculated solution of a single system and finally discuss error bounds for some special matrices.

In spite of the power of a high-speed computer, these theoretical relations discussed in the first part are of the utmost importance. We attempt to derive realistic (practical) bounds and we have opportunity in section 3 to show where the computed results must be very carefully examined; that is, they contradict our theoretical relations. Hence, our theoretical results can always be used at least in this negative way.

One of the concepts basic to a theoretical treatment of errors is that of a norm, and we feel it would be useful to give a brief discussion of this concept here, before proceeding to the body of the text. A full treatment of this topic can be found in *Computational Methods of Linear Algebra*, by Faddeev and Faddeeva. In dealing with vectors and matrices we frequently need some way of describing the concept of the size of a vector, x, or a matrix, A. We shall use the notation $N(x)$ or $N(A)$ to indicate this concept.

In two-dimensional Euclidean space, the size of a vector $x = (x_1, x_2)$ is just the length (always non-negative) of that vector:

$$N(x) = (x_1^2 + x_2^2)^{1/2}.$$

The length of such a vector has some immediate properties:

1. $N(x) > 0$ for $x \neq 0$ and $N(0) = 0$;
2. $N(cx) = |c|N(x)$ for any numerical multiplier c;
3. $N(x + y) \leq N(x) + N(y)$.

[1] Figures in brackets indicate the literature references on p. 309.

The usual length of a vector is just one way of describing its size. Other ways would depend upon our main point of emphasis. If we are interested in the distance in reference to the coordinate axes, we might take the norm of x to be the maximum distance from the origin along one of the axis:

$$N(x) = \max_i |x_i|.$$

For this definition, the same three properties also hold. The main thing, then, that we wish to associate with the concept of a norm is that it possess the three properties mentioned above. The generalization to n dimensions of each of the two examples given is obvious (although the proofs of some of the properties may not be), so let us consider the norm of a matrix.

DEFINITION: *The norm of a square matrix* A *is a nonnegative number* N(A) *which satisfies the conditions*

1. $N(A) > 0$, *if* $A \neq 0$ *and* $N(0) = 0$;
2. $N(cA) = |c|N(A)$;
3. $N(A + B) \leq N(A) + N(B)$
4. $N(AB) \leq N(A)N(B)$.

Some examples are in order. Similar to the length of a vector, we could take the square root of the sum of the squares of all the elements of A;

$$N(A) = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2} \qquad \text{(the Frobenius norm).}$$

Corresponding to the maximum element of a vector, we could take the maximum row sum of A:

$$N(A) = \max_i \sum_j |a_{ij}|.$$

Another one frequently used is also related to the maximum element of an $n \times n$ matrix:

$$N(A) = n \max_{i,j} |a_{ij}|.$$

In this case, the multiplier n is needed in order to satisfy the required properties.

Property 4 is peculiar to matrices and we have one special case to consider here; viz., the product Ax, a matrix times a vector. In this case we must choose the norms to be used in such a way that

$$N(Ax) \leq N(A)N(x).$$

This is called the compatibility condition. For each norm of a vector, there may be more than one norm of a matrix that can be used for $N(A)$ but not every matrix norm may satisfy this inequality.

Our main need for these norms arises when we consider the error of an approximate inverse X. We need some way to discuss the "size" of the error, $A^{-1} - X$, and the concept of a norm does this very well.

1. Error Bounds for the Approximate Inverse

In working with norms, frequent use if made of the property that $N(AB) \leq N(A)N(B)$. Before discussing various error bounds, it will be useful to examine this inequality, as it is the main reason why so many error estimates are overestimates. Let us take the Frobenius norm:

$$N(A) = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}.$$

Then, for a general A and B,

$$\begin{aligned}
N(AB)^2 &= \sum_{i,j} \left| \sum_k a_{ik} b_{kj} \right|^2 \\
&\leq \sum_{i,j} \left(\sum_k |a_{ik} b_{kj}| \right)^2 \\
&= \sum_{i,j,r,s} |a_{ir} b_{rj} a_{is} b_{sj}| \\
&= \sum_{i,j,r,s} |a_{ir} b_{sj}| \, |a_{is} b_{rj}| \\
&\leq \sum_{i,j,r,s} \frac{1}{2} \left(|a_{ir} b_{sj}|^2 + |a_{is} b_{rj}|^2 \right) \\
&= \sum_{i,j,r,s} |a_{ir} b_{sj}|^2 \\
&= \left(\sum_{i,r} |a_{ir}|^2 \right) \left(\sum_{s,j} |b_{sj}|^2 \right) \\
&= N(A)^2 N(B)^2.
\end{aligned}$$

From this derivation we can determine when equality is possible. If we consider matrices with only non-negative elements, the first inequality in the above becomes an equality. Hence, the remaining inequality results from the fact that $2ab \leq a^2 + b^2$ for all a and b. This is an equality if and only if $a = b$. In our problem, this means if and only if $a_{ir} b_{sj} = a_{is} b_{rj}$ or if and only if

$$\frac{a_{ir}}{a_{is}} = \frac{b_{rj}}{b_{sj}}$$

which says that the columns of A are in the same ratio as the rows of B. This conclusion can be written as $b_{rj} = (b_{sj}/a_{is}) a_{ir}$ for all r; that is, each column of B is a multiple of a row of A. The last statement can be seen very clearly if we examine the Cauchy-Bunyakovskii inequality which is the basis for the step in the above derivation that we are discussing. This inequality states that for two vectors, X and Y, we have $|(X, Y)| \leq |X| \cdot |Y|$ where (X, Y) is the scalar product. The familiar proof is as follows. Let us consider a nonzero X, and let $Z = Y - aX$, where $a = (Y, X)/(X, X)$. We note that $(Z, X) = (Y, X) - a(X, X) = 0$. Then

$$\begin{aligned}
|Z|^2 = (Z, Z) = (Z, Y - aX) = (Z, Y) = (Y - aX, \ Y) = (Y, Y) - a(X, Y) \\
= (Y, Y) - \frac{(Y, X)(X, Y)}{(X, X)} = \frac{|X|^2 |Y|^2 - |(X, Y)|^2}{|X|^2}.
\end{aligned}$$

Consequently $|X|^2 |Y|^2 - |(X, Y)|^2 = |X|^2 |Z|^2 \geq 0$, and the desired inequality follows. We will have equality if and only if $Z = 0$. This says that $Y = aX$; i.e., one vector is a scalar multiple of the other.

Although this knowledge of when equality occurs does not give us any idea of the magnitude of the inequality, it certainly indicates that $N(AB)$ will seldom be equal to $N(A)N(B)$. Since we have no knowledge of when the elements of A and B are going to be non-negative, the first inequality in the derivation just makes matters worse. If the signs of the elements of A and B are such that cancellation takes place, this inequality could also cause a tremendous difference. It has been our experience in general that the norm of a product is very much less than the product of the norms. (See the example at the end of this section.)

There is one case, however, when equality will always hold if N is the Frobenius norm. If U and V are unitary matrices, then $N(UAV) = N(A)$. To see this, we use $N(A)^2 = \text{trace } (AA^*)$, A^*

being the conjugate transpose of A, and the fact that if two matrices are similar, their traces are equal (since the trace is equal to the sum of the eigenvalues). Thus

$$\begin{aligned} N(UAV)^2 &= tr(UAVV^*A^*U^*) \\ &= tr(UAA^*U^*) \\ &= tr(AA^*) \\ &= N(A)^2. \end{aligned}$$

This equality has the following two consequences. If A is a normal matrix, it is unitarily equivalent to a diagonal matrix, and we have $N(AB) = N(U^*DUB) = N(DUB) \leq N(D)N(UB) = N(D)N(B) \leq n^{1/2} \max |l_i| N(B)$ where l_i are the eigenvalues of A. The second one also concerns the eigenvalues of A. We use a result of Schur which states that an arbitrary complex matrix A can be transformed into triangular form by a unitary matrix U. Now for any triangular matrix T, we have

$$\sum_i |t_{ii}|^2 \leq N(T)^2$$

with equality if and only if T is diagonal. Hence,

$$\sum_i |l_i|^2 = \sum_i |t_{ii}|^2 \leq N(T)^2 = N(U^*AU)^2 = N(A)^2$$

with equality if and only if A is unitarily equivalent to a diagonal matrix; i.e., when A is normal.

As a final note on the relation between $N(AB)$ and $N(A)N(B)$, we would like to mention that there exists no norm such that equality holds for all A and B. To prove this, we would just consider an $A = B$ such that $A \neq 0$ but $A^2 = 0$.

Our main concern will be with bounds for $N(A^{-1})$ and $N(A^{-1} - X)$ where X is an approximate inverse and N is any norm. To derive upper bounds for $N(A^{-1})$ is a nontrivial matter. Since A may be arbitrarily close to a singular matrix, uniform upper bounds for $N(A^{-1})$ are not to be found. Hence, we must make use of information that can be readily available. The most useful quantity for an approximate inverse X is the residual matrix $Y = I - AX$. The "size" of Y, of course, depends on how good an approximation X is but we have found that $N(Y)$ is invariably less than 1 and, in fact, actually less than $\frac{1}{2}$.

The fact that $N(Y)$ is computationally less than 1 turns out to be quite important because this inequality is basic to practically all our theoretical results. Its main use follows from the following two lemmas.

LEMMA 1.1. *If* B *is a matrix such that* $\mathrm{N(B)} < 1$, *then*

(a) $(\mathrm{I} \pm \mathrm{B})$ *is nonsingular and*

(b) $\mathrm{N}[(\mathrm{I} \pm \mathrm{B})^{-1}] \leq \dfrac{1}{1 - \mathrm{N(B)}}$ *if* $\mathrm{N(I)} = 1$.

PROOF: (a) Assume $(I \pm B)$ is a singular. Then there exists some vector $x \neq 0$ such that

$$(I \pm B)x = 0$$

or

$$x = \mp Bx$$

and

$$N(x) \leq N(B)N(x).$$

Since $N(x) > 0$, this last inequality means that $N(B) \geq 1$ which contradicts our condition that $N(B) < 1$.

(b) Since $(I \pm B)$ is nonsingular, we can let

$$R = (I \pm B)^{-1}.$$

Then

$$R^{-1} = (I \pm B)$$

$$I = R \pm RB$$

$$R = I \mp RB$$

and

$$N(R) \leq N(I) + N(R)N(B)$$

$$\leq 1 + N(R)N(B) \text{ if } N(I) = 1.$$

Then

$$N(R)[1 - N(B)] \leq 1$$

and

$$N(R) \leq \frac{1}{1 - N(B)}.$$

Another relation that needs the same condition and one that is also very useful is the following.

LEMMA 1.2. *If* B *is a matrix such that* N(B) < 1, *then*

$$(I - B)^{-1} = I + B + B^2 + \ldots + B^n + \ldots.$$

PROOF:

$$\text{Let } P_n = (I + B + B^2 + \ldots + B^{n-1})(I - B)$$

$$= I - B^n.$$

Then

$$I - P_n = B^n$$

and

$$N(I - P_n) = N(B^n)$$

$$\leq N(B)^n.$$

Since $N(B) < 1$, $N(B)^n$ approaches zero as n goes to infinity so that $N(I - P_n)$ approaches zero also and $I - P_n$ must converge to the 0 matrix, which means P_n converges to the identity matrix and hence $(I + B + B^2 + \ldots + B^{n-1})$ converges to $(I - B)^{-1}$.

With these two lemmas as background, we can proceed to derive the bounds for the matrices in which we are interested.

THEOREM 1.1. *Let* X *be an approximate inverse of* A *and let* Y = I − AX. *Then if* N(Y) < 1,

$$\frac{N(X)}{1 + N(Y)} \leq N(A^{-1}) \leq \frac{N(X)}{1 - N(Y)}.$$

PROOF: We shall prove the upper bound first.

$$Y = I - AX$$

$$AX = I - Y$$

$$X^{-1}A^{-1} = (I - Y)^{-1}$$

$$A^{-1} = X(I - Y)^{-1}$$

and

$$N(A^{-1}) \leq \frac{N(X)}{1 - N(Y)} \text{ if } N(I) = 1.$$

We may remove the restriction on $N(I)$ by the following:

$$A^{-1} = X(I - Y)^{-1}$$

$$= X(I + Y + Y^2 + Y^3 + \ldots)$$

$$= X + XY + XY^2 + XY^3 + \ldots$$

and

$$N(A^{-1}) \leq N(X) + N(X)N(Y) + N(X)N(Y^2) + \ldots$$
$$\leq N(X) + N(X)N(Y) + N(X)N(Y)^2 + \ldots$$
$$= N(X)[1 + N(Y) + N(Y)^2 + \ldots]$$
$$= \frac{N(X)}{1 - N(Y)}$$

For the lower bound, we proceed in a similar way.

$$Y = I - AX$$
$$AX = I - Y$$
$$X = A^{-1}(I - Y)$$
$$= A^{-1} - A^{-1}Y$$

and

$$N(X) \leq N(A^{-1})(1 + N(Y))$$

so that

$$\frac{N(X)}{1 + N(Y)} \leq N(A^{-1}).$$

We note that taking $Y = I - XA$, the lefthand residual matrix, would yield identical results.

COROLLARY. *For any nonsingular* A, *let* D *be the diagonal and let* $-$M *represent the off-diagonal part of* A.

Then, if $N(D^{-1}M) < 1$,

$$\frac{N(D^{-1})}{1 + N(D^{-1}M)} \leq N(A^{-1}) \leq \frac{N(D^{-1})}{1 - N(D^{-1}M)}.$$

PROOF: If we write $A = D - M = D(I - D^{-1}M)$, then $A^{-1} = (I - D^{-1}M)^{-1}D^{-1}$ and the proof follows as in the theorem.

Let us now consider the error matrix $A^{-1} - X$, and derive some bounds for its norm.

THEOREM 1.2. *Let* X *be an approximate inverse of* A *and let* Y = I $-$ AX. *Then if* N(Y) < 1,

$$N(A^{-1} - X) \leq \frac{N(XY)}{1 - N(Y)}.$$

PROOF:

$$Y = I - AX$$
$$AX = I - Y$$
$$X^{-1}A^{-1} = (I - Y)^{-1}$$
$$= I + Y + Y^2 + Y^3 + \ldots$$
$$A^{-1} = X + XY(I + Y + Y^2 + Y^3 + \ldots)$$
$$A^{-1} - X = XY(I - Y)^{-1}$$

so that

$$N(A^{-1} - X) \leq \frac{N(XY)}{1 - N(Y)},$$

where we have omitted the restriction that $N(I) = 1$ since it can be eliminated by the same process as used in the proof of Theorem 1.1.

COROLLARY. *Under the same conditions as Theorem 1.2,*

$$N(A^{-1}-X) \leq \frac{N(X)N(Y)}{1-N(Y)}.$$

PROOF: The proof follows immediately from the theorem and the properties of a norm.

We have developed this latter upper bound as a corollary since, as we mentioned at the beginning of this section, this bound is frequently very much larger than the one in the theorem. It is obviously simpler and demands very little additional computer time but the information may be misleading (see the example at the end of this section).

If we use the lefthand residual matrix, $Y_1 = I - XA$, the main difference from the above would be the need for $N(Y_1X)$ instead of $N(XY)$. However,

$$Y_1X = (I-XA)X = X - XAX = X(I-AX) = XY$$

so that actually the only difference comes from the use of $N(Y_1)$ instead of $N(Y)$. Unfortunately, it is not true in general that $Y_1 = Y$. In fact there are examples where there is a very large difference [4, p. 258]. So it is possible to have a good right inverse but a poor left inverse—good in the sense that the residual matrix is close to the zero matrix. Considering both residual matrices is of value, since they can give an indication of a poor inverse.

THEOREM 1.3. *Let* $E = (I - XA) - (I - AX)$. *Then*

$$N(A^{-1}-X) \geq \frac{N(E)}{2N(A)}.$$

PROOF:

$$\begin{aligned} E &= (I-XA) - (I-AX) \\ &= (A^{-1}-X)A - A(A^{-1}-X) \\ N(E) &\leq 2N(A)N(A^{-1}-X) \end{aligned}$$

so that

$$N(A^{-1}-X) \geq \frac{N(E)}{2N(A)}.$$

So the difference between the residual matrices gives us a lower bound which provides us with negative, but still useful, information.

The question of which residual matrix to use in Theorem 1.2 and/or its corollary still remains. Using the theorem and requiring that the norm of both residuals be less than one, there should be very little difference. In general, however, the fact that the norm of one residual is small does not guarantee that the norm of the other will also be small:

$$I - XA = A^{-1}(I-AX)A$$

so that

$$N(I-XA) \leq N(A)N(A^{-1})N(I-AX)$$

and similarly

$$N(I-AX) \leq N(A)N(A^{-1})N(I-XA).$$

For an ill-conditioned system, $N(A)N(A^{-1})$, the condition number, can be quite high. In practice, however, we find that the two residuals are almost invariably quite close, so that using Theorem 1.2 with either residual should yield satisfactory results. Since inverses are calculated mostly in such a way as to almost guarantee the smallness of $I-AX$, it would seem that using $I-XA$ in our error bounds would give more realistic information since we really expect a two-sided inverse. We also note that if it were particularly important to have $I-XA$ small, it would be safer to invert A^T, the transpose of A.

We have been considering only problems where the coefficient matrix had exact numerical entries. We would like to give now an error bound for the calculated inverse where there is some uncertainty in A [5]. Although we are working with the matrix A, we should really be working with A^*, which differs from A by an amount satisfying $N(A-A^*) \leq a$. We must, therefore, include this known difference in our bound for $N((A^*)^{-1}-X)$.

THEOREM 1.4. *Let* $\mathrm{N(A-A^*)} \leqslant \mathrm{a}$ *and let* $\mathrm{R = I - A^*X}$. *Then if* $\mathrm{N(R) < 1}$,

$$\mathrm{N((A^*)^{-1} - X) \leqslant \frac{N(X)[N(I - AX) + aN(X)]}{1 - N(I - AX) - aN(X)}}.$$

PROOF: We proceed exactly as in Theorem 1.2 and use R instead of Y, arriving at

$$N((A^*)^{-1} - X) \leqslant \frac{N(X)N(R)}{1 - N(R)}.$$

Now $R = I - A^*X = I - AX + AX - A^*X = (I - AX) + (A - A^*)X$ and $N(R) \leqslant N(I - AX) + N(A - A^*)N(X) \leqslant N(I - AX) + aN(X)$. Substituting this expression in the above inequality, we have the result stated in the theorem.

Let us now turn our attention to the relative error of the difference between A^{-1} and X.

THEOREM 1.5. *Let* X *be an approximate inverse of* A *and let* $\mathrm{Y = I - AX}$. *If* $\mathrm{N(Y) < 1}$,

$$\mathrm{\frac{1 - N(Y)}{N(X)} \leqslant \frac{1}{N(A^{-1})} \leqslant \frac{1 + N(Y)}{N(X)}}.$$

PROOF: The proof follows immediately from Theorem 1.1.

COROLLARY. $$\mathrm{1 - N(Y) \leqslant \frac{N(X)}{N(A^{-1})} \leqslant 1 + N(Y)}.$$

THEOREM 1.6. *Let* X *be an approximate inverse of* A *and let* $\mathrm{Y = I - AX}$. *If* $\mathrm{N(Y) < 1}$,

$$\mathrm{\frac{N(A^{-1} - X)}{N(A^{-1})} \leqslant \frac{N(XY)}{N(X)} \cdot \frac{1 + N(Y)}{1 - N(Y)}}.$$

PROOF: The proof follows directly from Theorem 1.2 and Theorem 1.5.

COROLLARY. $$\mathrm{\frac{N(A^{-1} - X)}{N(A^{-1})} \leqslant N(Y) \cdot \frac{1 + N(Y)}{1 - N(Y)}}$$

Actually, we have a simpler and better bound than this corollary.

THEOREM 1.7. *Let* X *be an approximate inverse of* A *and let* $\mathrm{Y = I - AX}$ *or* $\mathrm{I - XA}$. *Then*

$$\mathrm{\frac{N(A^{-1} - X)}{N(A^{-1})} \leqslant N(Y)}.$$

PROOF: $A^{-1} - X = A^{-1}(I - AX) = (I - XA)A^{-1}$ so that after taking norms, the inequaltiy follows.

This bound is not necessarily smaller than the one given in Theorem 1.6. If $N(Y)$ is small enough (say less than .1), the ratio in the above corollary can be expanded to the product

$$[1 + N(Y)][1 + N(Y) + N(Y)^2 + \ldots]$$

or to $1 + 2N(Y)$ if we ignore the higher powers of $N(Y)$. Then the corollary becomes

$$\frac{N(A^{-1} - X)}{N(A^{-1})} \leq N(Y)[1 + 2N(Y)] \approx N(Y),$$

again ignoring the higher powers of $N(Y)$. This indicates that in practice the bounds given in Theorem 1.7 and the corollary are quite close so that the bound given in Theorem 1.6 will probably still be the smaller one.

Finally, we also have a lower bound.

THEOREM 1.8. *Let* X *be an approximate inverse,* $Y = I - AX$, *and let* $E = (I - XA) - (I - AX)$. *Then if* $N(Y) < 1$,

$$\frac{N(A^{-1} - X)}{N(A^{-1})} \geq \frac{N(E)}{2N(A)} \cdot \frac{1 - N(Y)}{N(X)}.$$

PROOF: The proof follows immediately from Theorem 1.3 and Theorem 1.5.

We conclude this section with a numerical example of the above bounds. Let N be the maximum row sum norm and let A be the 6×6 segment of the infinite Hilbert matrix A, where $a_{ij} = 1/(i+j-1)$. We chose this particular matrix because of its notoriously poor condition number, about 2.92×10^8, but using only a 6×6 so that roundoff errors would not be too influential. Since we will use this matrix in all our examples, it might be good to give its *exact* inverse. For the nth order segment, if $A_n^{-1} = (b_{ij}^{(n)})$, then

$$b_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!}.$$

In our case (giving only the lower triangular part since it is symmetric) this turns out to be

$$A_6^{-1} = \begin{bmatrix} 36 & & & & & \\ -630 & 14700 & & & & \\ 3360 & -88200 & 564480 & & & \\ -7560 & 211680 & -1411200 & 3628800 & & \\ 7560 & -220500 & 1512000 & -3969000 & 4410000 & \\ -2772 & 83160 & -582120 & 1552320 & -1746360 & 698544 \end{bmatrix}$$

We note that this is not quite the inverse of the matrix actually used in the examples. It is impossible to represent all elements $1/(i+j-1)$ exactly in the computer and hence the input matrix is slightly different. For further information on this point of inexact data of the Hilbert matrix, see reference 12 (*Computer Solution of Linear Algebraic Systems* by Forsythe and Moler), section 19. We also indicate in section 6 how to avoid this problem.

To calculate X we ran the program given in the above text (p. 68). The computer we used was a Univac 1108 and we used double precision for all matrix products.

A word about notation. The standard computer notation $.235\,E+04$ means $.235 \times 10^4$ or 2350. The "E" means exponent to the base 10 and the "$+04$" is the positive exponent.

The absolute error: $(Y = I - AX)$

$$(1)\quad N(A^{-1} - X) \leq \frac{N(X)N(Y)}{1 - N(Y)} = .197\,E+06$$

$$(2)\quad N(A^{-1} - X) \leq \frac{N(XY)}{1 - N(Y)} = .612\,E-01$$

$$(3)\quad N(A^{-1}-X) \geqslant \frac{N(E)}{2N(A)} \qquad = .477\,E-02$$

The relative error:

$$(1')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant N(Y) \cdot \frac{1+N(Y)}{1-N(Y)} = .168\,E-01$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant N(Y) \qquad = .162\,E-01$$

$$(2')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant \frac{N(XY)}{N(X)} \cdot \frac{1+N(Y)}{1-N(Y)} = .522\,E-08$$

$$(3')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \geqslant \frac{N(E)}{2N(A)} \cdot \frac{1-N(Y)}{N(X)} = .393\,E-09$$

and for

$$\frac{N(X)}{1+N(Y)} \leqslant N(A^{-1}) \leqslant \frac{N(X)}{1-N(Y)},$$

$$.117E+08 \leqslant N(A^{-1}) \leqslant .121E+08$$

where $N(A^{-1}) = .11931731\,E+08$.

Note: If we had been able to input the data exactly, the norm of the inverse would have been .11865420 E+08.

It might be of interest to give the values of the previous quantities using the left hand residual, $I-XA$, for Y.

The absolute error:

$$(1)\quad N(A^{-1}-X) \leqslant \frac{N(X)N(Y)}{1-N(Y)} \qquad = .380\,E+06$$

$$(2)\quad N(A^{-1}-X) \leqslant \frac{N(YX)}{1-N(Y)} \qquad = .622\,E-01$$

$$(3)\quad N(A^{-1}-X) \geqslant \frac{N(E)}{2N(A)} \qquad = .477\,E-02$$

The relative error:

$$(1')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant N(Y)\,\frac{1+N(Y)}{1-N(Y)} \qquad = .328\,E-01$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant N(Y) \qquad = .309\,E-01$$

$$(2')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \leqslant \frac{N(YX)}{N(X)}\,\frac{1+N(Y)}{1-N(Y)} = .537\,E-08$$

$$(3')\quad \frac{N(A^{-1}-X)}{N(A^{-1})} \geqslant \frac{N(E)}{2N(A)}\,\frac{1-N(Y)}{N(X)} = .387\,E-09$$

and for

$$\frac{N(X)}{1+N(Y)} \leq N(A^{-1}) \leq \frac{N(X)}{1-N(Y)},$$

$$.116\,E+08 \leq N(A^{-1}) \leq .123\,E+08.$$

We also note:

$$N(XY) = .6024\,E-01$$

$$N(YX) = .6024\,E-01 \qquad \text{and} \qquad N(E) = .2337\,E-01.$$

It will also be of interest to give the values of the previous quantities using a different computer, the CDC 6400, with $Y=I-AX$. This machine uses about twice as many digits in its operations as the Univac 1108 and this is why the results are so much better. We found that the values of the elements of X were integers (to 8 digits).

The absolute error:

$$(1)\ N(A^{-1}-X) \leq \frac{N(X)N(Y)}{1-N(Y)} = .323\,E-01$$

$$(2)\ N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} = .178\,E-07$$

$$(3)\ N(A^{-1}-X) \geq \frac{N(E)}{2N(A)} = .555\,E-09$$

The relative error:

$$(1')\ \frac{N(A^{-1}-X)}{N(A^{-1})} \leq N(Y)\,\frac{1+N(Y)}{1-N(Y)} = .272\,E-08$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq N(Y) = .272\,E-08$$

$$(2')\ \frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{N(X)}\,\frac{1+N(Y)}{1-N(Y)} = .150\,E-14$$

$$(3')\ \frac{N(A^{-1}-X)}{N(A^{-1})} \geq \frac{N(E)}{2N(A)}\,\frac{1-N(Y)}{N(X)} = .468\,E-16$$

and for

$$\frac{N(X)}{1+N(Y)} \leq N(A^{-1}) \leq \frac{N(X)}{1-N(Y)},$$

$$.1186542\,E+08 \leq N(A^{-1}) \leq .1186542\,E+08.$$

We also note that $N(XY) = .178\,E-07$ and $N(E) = .272\,E-08$.

2. Error Bounds for the Approximate Solution Vector of $Ax=b$

In this chapter we shall derive estimates for the closeness of the calculated solution vector to the true solution vector of $Ax=b$. Our approach will be to assume that we have available both the approximate inverse X and its residual $Y=I-AX$ or $I-XA$. The necessity for this will be explained by the quotations at the end of this section.

Let z be the approximate solution vector of $Ax=b$ and let $r=b-Az$, the residual vector.

THEOREM 2.1. *The bound for the norm of the error is given by*

$$\mathrm{N}(\mathrm{A}^{-1}\mathrm{b}-\mathrm{z}) \leq \frac{\mathrm{N}(\mathrm{X})\mathrm{N}(\mathrm{r})}{1-\mathrm{N}(\mathrm{Y})}, \qquad \mathrm{N}(\mathrm{Y})<1.$$

PROOF:

$$b-Az=r$$

$$A^{-1}b-z=A^{-1}r$$

so that

$$N(A^{-1}b-z) \leq N(A^{-1})N(r)$$

$$\leq \frac{N(X)N(r)}{1-N(Y)}$$

using our estimates for $N(A^{-1})$ from the previous section.

THEOREM 2.2. *Let* $\mathrm{z}=\mathrm{Xb}$. *Then if* $\mathrm{N}(\mathrm{Y})<1$,

$$N(A^{-1}b-z) \leq \frac{N(XY)N(b)}{1-N(Y)}.$$

PROOF:

$$A^{-1}b-z=A^{-1}b-Xb$$

$$=(A^{-1}-X)b$$

and

$$N(A^{-1}b-z) \leq \frac{N(XY)N(b)}{1-N(Y)}$$

again using information from the previous section.

COROLLARY. $N(A^{-1}b-z) \leq \dfrac{N(X)N(Y)N(b)}{1-N(Y)}$.

We would like to make a remark stemming from the proof of Theorem 2.1. If the equality $A^{-1}r=A^{-1}b-z$ is ever used in practice, say for checking a process or computer program where the solution is known, we should note that we would be using X and not A^{-1}. That is to say

$$A^{-1}r=A^{-1}(b-Az)=A^{-1}b-A^{-1}Az=A^{-1}b-Iz=A^{-1}b-z$$

but

$$Xr=X(b-Az)=Xb-XAz=Xb-(I-Y_1)z=Xb-z+Y_1z,$$

where Y_1 is the lefthand residual $I-XA$. Hence, we have an additional factor to keep in mind, Y_1z. The need for this obviously depends on the quality of X; i.e., the closeness of XA to I. There are examples, however, where X may be a good inverse but where Xb may not be a good approximation to the solution vector [1, p. 149]. So, in general, the correction factor Y_1z can be of impor-

406-680 O - 71 - 3

tance. To give some indication of just how important, let us give an example.

In solving $Ax=b$ where A was the 6×6 Hilbert segment and b was $(1,2,3,4,5,6)^T$, we used Gaussian Elimination (with iterative improvement) to get X and the machine was an IBM 7094. For the maximum element norm, $N(Xr)$ was of order 10^{-2}; $N(Y_1z)$, of order 10^{-1}; $N(Xb-z)$, of order 10^{-1}. The need for the correction factor is clear. The fact that the Hilbert matrix is ill-conditioned results in the high order of the norms but the need is still there to get equality even for a well-conditioned matrix. Using a diagonally dominant tridiagonal matrix for A, we found that $N(X_r)$ was of order 10^{-7}; $N(Y_1z)$, of order 10^{-8}; $N(Xb-z)$, of order 10^{-8}. The need for the correction factor is again clear although not so vital. We would also like to emphasize that the residual to be used is the left hand one. As shown in section 1, there can be an appreciable difference in the two.

The use of the lefthand residual (only) also gives us a bound for the relative error.

THEOREM 2.3. *Let* $z=Xb$. *Then*

$$\frac{N(A^{-1}b-z)}{N(A^{-1}b)} \leq N(I-XA).$$

PROOF:
$$A^{-1}b-z=A^{-1}b-Xb$$
$$=(A^{-1}-X)b$$
$$=(I-XA)A^{-1}b$$

so that
$$\frac{N(A^{-1}b-z)}{N(A^{-1}b)} \leq N(I-XA).$$

Once again this residual seems to be of more practical use than $I-AX$. (It was noted in the previous section that $I-AX$ is ordinarily small by reason of the numerical procedure so that checking $N(I-XA)$ is the important thing.) We can see that a small right residual $I-AX$ may not guarantee the smallness of the error vector from the following:

$$A^{-1}b-z=A^{-1}b-Xb=(A^{-1}-X)b=A^{-1}(I-AX)b.$$

Hence, if A^{-1} has a large norm, the norm of the error vector may not be small.

It can also be seen from the proof of Theorem 2.1 that a small residual vector may *not* indicate a good solution vector:

$$A^{-1}b-z=A^{-1}r$$

so that
$$N(A^{-1}b-z) \leq N(A^{-1})N(r).$$

Again, if $N(A^{-1})$ is large, $N(A^{-1}b-z)$ could be large also. On the other hand, a large residual vector can indicate a poor solution vector. We show this in the following theorem.

THEOREM 2.4. *If* $r=b-Az$, *then*

$$N(A^{-1}b-z) \geq \frac{N(r)}{N(A)}.$$

PROOF:
$$r=b-Az$$
$$=A(A^{-1}b-z)$$
$$\frac{N(r)}{N(A)} \leq N(A^{-1}b-z).$$

We would like to include here a brief discussion of an error bound for the computed solution vector of a system that does not have exact data [5, p. 151]. As in Theorem 1.4, let

$$N(A-A^*) \leq a,\ N(b-b^*) \leq c,\ x^* = A^{-1}b,\ x^{**} = (A^*)^{-1}b^* \text{ and } z = Xb,$$

where X is the approximate inverse of A.

THEOREM 2.5. *Let* $R = I - A^*X$. *Then if* $N(R) < 1$,

$$N(z-x^{**}) \leq \frac{N(X)[aN(z)+N(b-Az)+c]}{1-N(I-AX)-aN(X)}.$$

PROOF:
$$z-x^{**} = z-(A^*)^{-1}b^*$$
$$= (A^*)^{-1}(A^*z-b^*)$$

and
$$N(z-x^{**}) \leq N[(A^*)^{-1}]N(A^*z-b^*).$$

Now
$$A^*z-b^* = A^*z-Az+Az-b+b-b^*$$

so that
$$N(A^*z-b^*) \leq N(A^*-A)N(z)+N(Az-b)+N(b-b^*).$$

To get a bound for $N[(A^*)^{-1}]$ we proceed as in section 1, starting with $R = I-A^*X$, which leads to

$$N[(A^*)^{-1}] \leq \frac{N(X)}{1-N(R)}.$$

Also
$$R = I-A^*X$$
$$= I-AX+AX-A^*X$$

and
$$N(R) \leq N(I-AX)+N(A-A^*)N(X).$$

Combining all of the above, we have the result stated in the theorem.

As mentioned at the beginning of the section, we have used the approximate inverse in all our bounds. This is most unfortunate because it means that we have to solve an additional n systems to get information about the solution of one system. It seems, however, that to expect any improvement in this situation is rather futile, according to some of the best men in the field. We shall quote from a few of them:

A. M. Turing: (1948) "It is difficult to determine the accuracy of the solution of a set of equations without inverting the matrix." [6, p. 301]

J. H. Wilkinson: (1963) "It does not seem possible to obtain a rigorous estimate of the accuracy of a solution to a set of equations without some estimate for $N(A^{-1})$. . . . To achieve more . . . we need an approximate inverse and a bound for its error." [7, p. 126]

E. L. Albasiny: (1964) "Unless one has some knowledge of the inverse of A we cannot immediately tell however how accurate our solution is." [8, vol. 1, p. 171]

L. Fox: (1965) "The search for the worst combination of uncertainties involves the computation of the inverse, and in solving linear equations we need a simpler and reasonably satisfactory criterion for the number of quotable figures in the calculated solution. If the coefficients vary widely in magnitude there is unlikely to be any satisfactory analysis . . ." [1, p. 159]

A. S. Householder: (1965)	"For error estimates it is advantageous to find at least an approximation to A^{-1} even when only a single vector x is required." [9, p. 91]
W. Kahan: (1966)	"The only disadvantage that can be occasioned by the lack of an estimate Z of A^{-1} is that there is no other way to get a rigorous errorbound for $z-x$." [10, p. 785]
C. Moler: (1967)	". . . without a bound for the condition of A, no precise conclusion can be made regarding $N(x_n-x)$." [11, p. 320]

Hence, it would seem that the bounds given in this section are about the only ones possible.

In some special cases, however, where A^{-1} is easily obtained or estimated, some improvement in this situation might be hoped for. We will discuss some of these in section 5.

3. Improvement Schemes for the Approximate Inverse

In this section we will present three schemes for improving the inverse that we have found quite effective. Each has a different order of convergence but not without some expense. We then generalize to a scheme of any order of convergence. Next, we show how two iteration methods which are frequently used are special cases of our schemes. Finally we give a numerical example of the various bounds derived.

The first scheme follows directly from the same procedure that yielded our error bounds for the approximate inverse in section 1. Let X be the approximate inverse and let $Y=I-AX$. Then

$$A^{-1}=X+XY+XY^2+\ldots.$$

The righthand side of this equality gives us our first scheme:

$$X_k=X+XY+XY^2+\ldots+XY^k, \qquad k\geq 0.$$

If $N(Y)<1$, we have convergence as k goes to infinity and by taking enough terms we can get as close to the true inverse as we please. The scheme is ideal for computer execution, and double precision (at least) should be used for the products of the matrices and also for the sum. The result would actually be a double precision inverse and we have found this inverse to be very effective in yielding $A^{-1}b$, the actual solution vector for a linear algebraic system.

The same scheme can give us error bounds for the closeness of X to A^{-1}. We have

$$A^{-1}-X=XY+XY^2+\ldots$$

so that we can get an excellent estimate for the difference just by taking enough terms. Hence, given any X we can tell how good it is, the only requirement being that $N(Y)<1$. If that particular X is not good enough, we can take more terms from the series until we are satisfied:

$$\begin{aligned} A^{-1}-X_k&=XY^{k+1}+XY^{k+2}+\ldots \\ &=XY^{k+1}(I+Y+Y^2+\ldots) \end{aligned}$$

so that

$$\begin{aligned} N(A^{-1}-X_k)&\leq\frac{N(XY^{k+1})}{1-N(Y)} \\ &\leq\frac{N(X)N(Y^{k+1})}{1-N(Y)} \\ &\leq\frac{N(X)N(Y)^{k+1}}{1-N(Y)}. \end{aligned}$$

As mentioned previously, the first bound is by far the best. However, the last bound can be quite useful. Having calculated X and Y, we can easily determine k to satisfy our requirements on

$N(A^{-1}-X)$ (realizing that this is definitely an overestimate), and proceed to sum the correct number of terms from our series and also determine the kind of multiprecision that may be needed to do the calculation correctly. Using this last bound also eliminates one matrix multiplication, to get XY^{k+1}.

If we use the lefthand residual, the procedure is essentially the same and we have for

$$N(I-XA)=N(Y)<1,$$

$$A^{-1}=X+YX+Y^2X+\ldots$$

$$X_k=X+YX+Y^2X+\ldots+Y^kX$$

and

$$N(A^{-1}-X_k)\leqslant\frac{N(Y^{k+1}X)}{1-N(Y)}$$

$$\leqslant\frac{N(X)N(Y^{k+1})}{1-N(Y)}$$

$$\leqslant\frac{N(X)N(Y)^{k+1}}{1-N(Y)}.$$

The second scheme we are interested in is very similar but the residual matrix must be recalculated at each step. We start as before but now let $X_0=X$ and define a recursion formula:

$$X_k=X_{k-1}+XY^k=X+XY+XY^2+\ldots+XY^k$$
$$Y_k=I-AX_k,\qquad k>0.$$

If $N(Y)<1$, we are assured of convergence as k goes to infinity and with Y_k as defined we proceed exactly as in section 1 to get

$$N(A^{-1}-X_k)\leqslant\frac{N(X_kY_k)}{1-N(Y_k)},\qquad N(Y_k)<1.$$

In our derivation for this bound we assumed X_k^{-1} exists. It is clear that one of our basic assumptions from the beginning has been that X^{-1} exists. It seems, however, that we should be more explicit about the existence of X_k^{-1}.

$$X_k=X+XY+XY^2+\ldots+XY^k$$
$$=X(I+Y+Y^2+\ldots+Y^k).$$

Now

$$(I-Y)^{-1}=I+Y+Y^2+\ldots+Y^k+Y^{k+1}+\ldots.$$

Hence

$$I+Y+Y^2+\ldots+Y^k=(I-Y)^{-1}-Y^{k+1}(I+Y+Y^2+\ldots)$$

$$=(I-Y)^{-1}-Y^{k+1}(I-Y)^{-1}$$

$$=(I-Y^{k+1})(I-Y)^{-1}.$$

Since $N(Y^{k+1})\leqslant N(Y)^{k+1}\leqslant N(Y)<1$, $(I-Y^{k+1})^{-1}$ exists and hence we are assured of the existence of X_k^{-1}.

In this scheme the residual actually gets multiplied by itself as the process continues:

$$\begin{aligned} Y_1 &= I - AX_1 \\ &= I - A(X + XY) \\ &= Y - AXY \\ &= (I - AX)Y \\ &= Y^2. \end{aligned}$$

and

$$\begin{aligned} Y_k &= I - AX_k \\ &= I - A(X_{k-1} + XY^k) \\ &= Y_{k-1} - AXY^k \\ &= Y^k - AXY^k \quad \text{(by induction)} \\ &= Y^{k+1}. \end{aligned}$$

Since $N(Y_k) \leq N(Y)^{k+1} < N(Y)$, our condition for convergence and the validity of the error bound is still $N(Y) < 1$.

Now $$X_k = X + XY + XY^2 + \ldots + XY^k$$

so that $$N(X_k) \leq N(X)[1 + N(Y) + N(Y)^2 + \ldots + N(Y)^k].$$

Since $$\frac{1 - p^{k+1}}{1 - p} = 1 + p + p^2 + \ldots + p^k, \qquad p \neq 1,$$

we have $$N(X_k) \leq N(X) \cdot \frac{1 - N(Y)^{k+1}}{1 - N(Y)}.$$

We also have $$\begin{aligned} X_k Y_k &= (X + XY + XY^2 + \ldots + XY^k)Y^{k+1} \\ &= XY^{k+1} + XY^{k+2} + \ldots + XY^{2k+1} \end{aligned}$$

so that $$\begin{aligned} N(X_k Y_k) &\leq N(XY^{k+1})[1 + N(Y) + \ldots + N(Y)^k] \\ &= N(XY^{k+1}) \frac{1 - N(Y)^{k+1}}{1 - N(Y)}. \end{aligned}$$

Hence $$\begin{aligned} \frac{N(X_k Y_k)}{1 - N(Y_k)} &\leq \frac{N(X_k Y_k)}{1 - N(Y)^{k+1}} \\ &\leq \frac{N(XY^{k+1})}{1 - N(Y)} \end{aligned}$$

and the best bound for our second scheme is better than that for our first.

Summarizing the bounds for the second scheme, we have for $N(Y) < 1$,

$$N(A^{-1}-X_k) \leqslant \frac{N(X_kY_k)}{1-N(Y_k)}$$

$$\leqslant \frac{N(X_k)N(Y)^{k+1}}{1-N(Y)^{k+1}}$$

$$\leqslant \frac{N(X)N(Y)^{k+1}}{1-N(Y)}.$$

It will be noticed that the last bound is exactly the same as for the first scheme, so that any improvement must come by using the first bound.

In spite of the fact that we recalculate the residual matrix at each step, we used it only in the error bound expression. It would seem useful to try to get it involved in the calculation of the new approximate inverse also. We can do this quite simply and very effectively by returning to the basic relation at each step. The process would be as follows:

$$A^{-1}=X+XY+XY^2+\ldots$$

with convergence if $N(Y) < 1$.

$$X_1=X+XY=X(I+Y)$$

$$Y_1=I-AX_1=Y^2.$$

Using this expression for Y_1, we repeat:

$$A^{-1}=X_1+X_1Y_1+X_1Y_1^2+\ldots$$

with convergence if $N(Y_1) < 1$

but $N(Y_1) \leqslant N(Y)^2$.

$$X_2=X_1+X_1Y_1=X_1(I+Y_1)$$
$$Y_2=I-AX_2=Y_1^2=Y^4=Y^{2^2}.$$

In general we have

$$X_{r+1}=X_r(I+Y_r)$$
$$Y_{r+1}=I-AX_{r+1}.$$

Now

$$\begin{aligned} Y_{r+1} &= I-A(X_r(I+Y_r)) \\ &= I-AX_r-AX_rY_r \\ &= Y_r-AX_rY_r \\ &= Y_r^2 \end{aligned}$$

so that

$$Y_r=Y^{2^r}.$$

Hence, if $N(Y) < 1$ we have convergence for our third scheme and since the error matrix is squared at each step, we have quadratic convergence.

It is interesting to observe that this third scheme actually amounts to taking a very large number of terms in our first series at each step. For example,

$$\begin{aligned} X_3 &= X_2(I+Y_2) \\ &= X_1(I+Y_1)(I+Y_1^2) \\ &= X(I+Y)(I+Y^2)(I+Y^{2^2}) \\ &= X+XY+XY^2+\ldots+XY^7 \end{aligned}$$

and in general,

$$\begin{aligned} X_r &= X(I+Y)(I+Y^2)\ \ldots\ (I+Y^{2^{r-1}}) \\ &= X+XY+XY^2+\ldots+XY^{2^r-1}. \end{aligned}$$

The recurrence formula, however, is much simpler and more efficient.

The bounds for the third scheme follow directly as before. If $N(Y)<1$,

$$\begin{aligned} N(A^{-1}-X_r) &\leqslant \frac{N(X_rY_r)}{1-N(Y_r)} \\ &\leqslant \frac{N(X_r)N(Y_r)}{1-N(Y_r)} \\ &\leqslant \frac{N(X_r)N(Y)^{2^r}}{1-N(Y)^{2^r}} \\ &\leqslant \frac{N(X)N(Y)^{2^r}}{1-N(Y)}. \end{aligned}$$

The existence of X_r^{-1} can be shown in the same way as in our second scheme.

If we had used the lefthand residual matrix, the obvious changes would follow as before and we also have $Y_rX_r=X_rY_r$.

We can generalize our approach in the sense that we can define a recursion such that the residual (error) matrix can be any power of the previous residual matrix. For example, to get third-order convergence, we take the first three terms of our basic series and repeat after each step. Thus

$$\begin{aligned} X_{s+1} &= X_s(I+Y_s+Y_s^2) \\ Y_{s+1} &= I-AX_{s+1} \end{aligned}$$

with $X_0=X$, $Y_0=Y=I-AX$ and $N(Y)<1$. The proofs follow exactly as in the second-order case and we have

$$Y_{s+1}=Y^{3^{s+1}}.$$

However, another matrix multiplication needed at each step and the possible loss of significant figures are factors to be considered in the practical application of this scheme.

For any order convergence, say t, we have

$$\begin{aligned} X_{s+1} &= X_s(I+Y_s+Y_s^2+\ldots+Y_s^{t-1}) \\ Y_{s+1} &= I-AX_{s+1}. \end{aligned}$$

The proof is straightforward by induction on t and we also have

$$Y_{s+1}=Y^{t^{s+1}}.$$

The error bounds follow directly as before:
If $N(Y) < 1$,

$$\begin{aligned} N(A^{-1}-X_s) &\leq \frac{N(X_sY_s)}{1-N(Y_s)} \\ &\leq \frac{N(X_s)N(Y_s)}{1-N(Y_s)} \\ &\leq \frac{N(X_s)N(Y)^{t^s}}{1-N(Y)^{t^s}} \\ &\leq \frac{N(X)N(Y)^{t^s}}{1-N(Y)}, \end{aligned}$$

the first being the best, but the last being useful in determining s as was indicated on page 266.

We note that the multiprecision needed for the effective use of these higher order convergence schemes might be a minimal problem in view of the hardware being developed by the computer industry. At the end of section 1 we gave an example using different computers and we found that the CDC 6400 gave extraordinary results.

Let us now discuss the relative error of the approximate inverse derivable from these schemes. For each scheme we can use

$$A^{-1}-X_k = A^{-1}(I-AX_k) = (I-X_kA)A^{-1}$$

so that

$$\frac{N(A^{-1}-X_k)}{N(A^{-1})} \leq N(Y_k),$$

where Y_k may represent the lefthand or the righthand residual matrix. Hence, for our first scheme:

$$\frac{N(A^{-1}-X_k)}{N(A^{-1})} \leq N(Y);$$

for the second:

$$\begin{aligned} \frac{N(A^{-1}-X_k)}{N(A^{-1})} &\leq N(Y^{k+1}) \\ &\leq N(Y)^{k+1}; \end{aligned}$$

and for the third:

$$\begin{aligned} \frac{N(A^{-1}-X_r)}{N(A^{-1})} &\leq N(Y^{2^r}) \\ &\leq N(Y)^{2^r}. \end{aligned}$$

In general for tth order convergence:

$$\begin{aligned} \frac{N(A^{-1}-X_s)}{N(A^{-1})} &\leq N(Y^{t^s}) \\ &\leq N(Y)^{t^s}. \end{aligned}$$

Hence, if we use this relative error as a criterion for stopping the process, it is evident that the third is far better than the first two. There are also other considerations which will be mentioned shortly. To get a better bound we can use the relation

$$\frac{1}{N(A^{-1})} \leq \frac{1+N(Y_k)}{N(X_k)},$$

which follows exactly as in section 1. Hence, for our first scheme:

$$\frac{N(A^{-1}-X_k)}{N(A^{-1})} \leqslant \frac{N(XY^{k+1})}{N(X)} \cdot \frac{1+N(Y)}{1-N(Y)}$$

$$\leqslant N(Y)^{k+1} \cdot \frac{1+N(Y)}{1-N(Y)};$$

for the second scheme:

$$\frac{N(A^{-1}-X_k)}{N(A^{-1})} \leqslant \frac{N(X_kY_k)}{N(X_k)} \cdot \frac{1+N(Y_k)}{1-N(Y_k)}$$

$$\leqslant N(Y_k) \cdot \frac{1+N(Y_k)}{1-N(Y_k)}$$

$$\leqslant N(Y)^{k+1} \cdot \frac{1+N(Y)^{k+1}}{1-N(Y)^{k+1}};$$

and in general for the $t$the order convergence scheme:

$$\frac{N(A^{-1}-X_s)}{N(A^{-1})} \leqslant \frac{N(X_sY_s)}{N(X_s)} \cdot \frac{1+N(Y_s)}{1-N(Y_s)}$$

$$\leqslant N(Y_s) \cdot \frac{1+N(Y_s)}{1-N(Y_s)}$$

$$\leqslant N(Y)^{t^s} \cdot \frac{1+N(Y)^{t^s}}{1-N(Y)^{t^s}}.$$

Let us compare the three main schemes we have discussed. Our basis for comparison will be the number of new matrix multiplications needed to advance one step in the process and get the error bound for that particular iterate, presuming the previous one was not small enough. This means, of course, that the error bound must be calculated at each step. We have written the third scheme as $X_k+X_kY_k$ instead of $X_k(I+Y_k)$ since the product X_kY_k was calculated in the previous step for the evaluation of the error bound and is available. We will identify the first, second, and third schemes by Scheme I, Scheme II, and Scheme III in the following table and list the quantities that must be calculated in going from one step to the next.

	Scheme I		Scheme II		Scheme III	
	Step k	Step $k+1$	Step k	Step $k+1$	Step k	Step $k+1$
Iterate.....	X_k	X_k+XY^{k+1}	X_k	X_k+XY^{k+1}	X_k	$X_k+X_kY_k$
Residual.....	Y	Y	Y_k	Y_kY	Y_k	Y_kY_k
Bound for $N(A^{-1}-X_k)$...	$\frac{N(XY^{k+1})}{1-N(Y)}$	$\frac{N(XY^{k+2})}{1-N(Y)}$	$\frac{N(X_kY_k)}{1-N(Y_k)}$	$\frac{N(X_{k+1}Y_{k+1})}{1-N(Y_{k+1})}$	$\frac{N(X_kY_k)}{1-N(Y_k)}$	$\frac{N(X_{k+1}Y_{k+1})}{1-N(Y_{k+1})}$
Additional products.....	1 XY^{k+2}		3 XY^{k+1}, Y_kY, $X_{k+1}Y_{k+1}$		2 Y_kY_k, $X_{k+1}Y_{k+1}$	

We see that the second scheme has one more matrix multiplication than the third and, since the latter has a much better rate of convergence, the third scheme is definitely better than the second. In comparing the first and the third, the question of the additional matrix multiplication to get better convergence arises and cannot be answered in general. The exponent of Y in Scheme I may be such as to demand multiprecision hardware and this may be worse than performing an

extra matrix multiplication. In some cases too, this extra multiplication may itself demand multiprecision hardware. In view of the quadratic convergence, though, many fewer steps should be required for the third scheme and probably multiprecision hardware would not be needed (beyond double precision, that is). We note that in our second scheme, the convergence is better than in the first and we have proven that the bounds are also better. *Two* additional matrix multiplications may not be worth it, however. Considering all of these factors and also keeping in mind the relative error, we feel that the third scheme is the most practical.

We would like to show now that two frequently used iteration procedures to find the inverse of a matrix are actually included in our development. The first is defined by the recurrence

$$X_{k+1}=X+(I-XA)X_k, \qquad k\geq 0, \tag{A}$$

with $X_0=0$. Letting $Y_k=I-AX_k$, we can rewrite this recurrence as

$$\begin{aligned} X_{k+1}&=X+X_k-XAX_k\\ &=X_k+X(I-AX_k)\\ &=X_k+XY_k. \end{aligned}$$

Now

$$\begin{aligned} Y_k&=I-AX_k\\ &=I-A(X_{k-1}+XY_{k-1})\\ &=I-AX_{k-1}-AXY_{k-1}\\ &=(I-AX)Y_{k-1}\\ &=YY_{k-1} \end{aligned}$$

and by induction we have

$$Y_k=Y^k.$$

So this recurrence can be written

$$\begin{aligned} X_{k+1}&=X_k+XY^k, \qquad k\geq 0\\ Y_{k+1}&=I-AX_{k+1}, \end{aligned}$$

which is exactly our second scheme (the starting value is the only difference). However, in the original form, only the lefthand residual can be used whereas in our development either one is applicable. We note also that for the convergence of (A) each eigenvalue of the righthand residual must be of modulus less than one. Although this seems to imply the necessity of working with both residuals, we realize that this is not the case since the two residuals are similar:

$$I-AX=A(I-XA)A^{-1}.$$

The second frequently used iteration is defined by

$$X_{k+1}=X_k(2I-AX_k), \qquad k\geq 0 \tag{B}$$

where X_0 is arbitrary. The immediate difficulty is the choice of X_0 and hence this iteration is more useful as an improvement scheme. If we let $X_0=X$ and $Y_k=I-AX_k$, we can rewrite (B) as

$$X_{k+1}=X_k(I+Y_k)$$

which is identical to our third scheme. The changes to make use of the lefthand residual, once (B) is written this way, follow exactly as in our development.

So we see that both of these iteration formulas can be thought of as having their foundation in our basic relation that

$$A^{-1}=X+XY+XY^2+\ldots.$$

We have repeatedly mentioned that in our schemes we can improve either the lefthand or the righthand inverse. This is a very important consideration in the solution of $Ax=b$, as will be pointed out in the next section.

For a numerical example, let A be the 6×6 segment of the Hilbert Matrix and to generate the approximate inverse, X, we use the same program and computer as in section 1, except that we used no iterative improvement scheme. We used double precision accumulation for the products XY and our criterion for stopping was when the elements of XY^k were in magnitude less than 10^{-8}. The final results for each scheme were identical.

For convenience we list the results separately and then make some comments about them.

Scheme I

$$N(A^{-1}-X)\leq N(XY+XY^2+\ldots+XY^5) \qquad =.412\,E+05$$

$$N(A^{-1}-X_5)\leq\frac{N(XY^6)}{1-N(Y)} \qquad =.227\,E-07$$

$$\leq\frac{N(X)N(Y)^6}{1-N(Y)} \qquad =.199\,E-02$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})}\leq N(Y) \qquad =.234\,E-01$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})}\leq N(XY+XY^2+\ldots+XY^5)*\frac{1+N(Y)}{N(X)} \qquad =.360\,E-02$$

$$\frac{N(A^{-1}-X_5)}{N(A^{-1})}\leq\frac{N(XY^6)}{N(X)}\,\frac{1+N(Y)}{1-N(Y)} \qquad =.191\,E-14$$

$$\leq N(Y)^6\frac{1+N(Y)}{1-N(Y)} \qquad =.170\,E-09$$

and for

$$\frac{N(X_5)}{1+N(Y)}\leq N(A^{-1})\leq\frac{N(X_5)}{1-N(Y)},$$

$$.117\,E+08\leq N(A^{-1})\leq .122\,E+08$$

Scheme II

$$N(A^{-1}-X_5) \leqslant \frac{N(X_5Y_5)}{1-N(Y_5)} = .114\ E-04$$

$$\leqslant \frac{N(X_5)N(Y_5)}{1-N(Y_5)} = .502\ E-04$$

$$\leqslant \frac{N(X)N(Y)^6}{1-N(Y)^6} = .194\ E-02$$

$$\frac{N(A^{-1}-X_5)}{N(A^{-1})} \leqslant N(Y_5) = .421\ E-11$$

$$\frac{N(A^{-1}-X_5)}{N(A^{-1})} \leqslant \frac{N(X_5Y_5)}{N(X_5)} \cdot \frac{1+N(Y_5)}{1-N(Y_5)} = .952\ E-12$$

$$\leqslant N(Y)^6 \cdot \frac{1+N(Y)^6}{1-N(Y)^6} = .162\ E-09$$

and for

$$\frac{N(X_5)}{1+N(Y_5)} \leqslant N(A^{-1}) \leqslant \frac{N(X_5)}{1-N(Y_5)},$$

$$.11931731\ E+08 \leqslant N(A^{-1}) \leqslant .11931731\ E+08$$

Scheme III

$$N(A^{-1}-X_3) \leqslant \frac{N(X_3Y_3)}{1-N(Y_3)} = .840\ E-05$$

$$\leqslant \frac{N(X_3)N(Y_3)}{1-N(Y_3)} = .505\ E-04$$

$$\leqslant \frac{N(X)N(Y)^8}{1-N(Y)} = .108\ E-05$$

$$\frac{N(A^{-1}-X_3)}{N(A^{-1})} \leqslant N(Y_3) = .424\ E-11$$

$$\frac{N(A^{-1}-X_3)}{N(A^{-1})} \leqslant \frac{N(X_3Y_3)}{N(X_3)} \cdot \frac{1+N(Y_3)}{1-N(Y_3)} = .704\ E-12$$

$$\leqslant N(Y)^8 \cdot \frac{1+N(Y)^8}{1-N(Y)^8} = .886\ E-13$$

and

$$\frac{N(X_3)}{1+N(Y_3)} \leqslant N(A^{-1}) \leqslant \frac{N(X_3)}{1-N(Y_3)},$$

$$.11931731\ E+08 \leqslant N(A^{-1}) \leqslant .11931731\ E+08.$$

We would like to make some comments about this example. One of our theoretical results was that

$$N(Y_k) \leq N(Y)^{k+1}.$$

Now $N(Y)^6 \approx .162\ E-09$, so that this inequality does actually hold.

Another inequality we proved was that for Scheme II

$$\frac{N(X_kY_k)}{1-N(Y_k)} \leq \frac{N(XY^{k+1})}{1-N(Y)}.$$

We notice here, however, that this is not the case for $k=5$. In investigating further we found that this inequality did hold for $k=1, 2, 3$, and for $k=4$ there was almost equality. In examining the significant digits involved, we found that after $k=3$, at which point $N(Y_3) \approx .1\ E-08$, we had gone beyond the accuracy of the numbers with which we were dealing. We were working in double precision with numbers originally of the order $E+08$ and have now gone beyond 16 digits. Hence, any information beyond this point becomes meaningless.

The difficulty is more apparent in dealing with X_kY_k and XY^{k+1} than with just Y_k and Y because of the matrix multiplications involved which lead to more serious round off errors. This is the reason that we mentioned in this chapter the possible need for multiprecision when working with these schemes.

The contradiction in Scheme III is different but comes from the same source. Here

$$\frac{N(X_3Y_3)}{1-N(Y_3)} > \frac{N(X)N(Y)^8}{1-N(Y)}$$

and $k=3$, the quadratic convergence making the difficulty show up earlier.

This example very clearly illustrates the point of the question asked in the Introduction: How do we know these numbers *are* the solution of our problem? The bounds developed here certainly help in answering this question.

4. Improvement Schemes for the Approximate Solution Vector of $Ax=b$

Let us now examine ways of improving the solution vector, z, of a linear algebraic system, $Az=b$. As in the previous section, the residual, $r=b-Az$, plays an important role. The standard approach is to solve the system $Ax=r$ whose solution, d, is an improvement to be added to z. The usual convergence proof depends on Wilkinson's backward analysis which says that z is the exact solution of a slightly perturbed system $(A+E)x=b$. The condition for convergence of the iterative improvement scheme [12] is that $N(A^{-1}E) < \frac{1}{2}$. We would like to present a slightly different scheme and give two convergence proofs for it. We will then give a different convergence proof for the standard approach, one we find more meaningful.

The alternate scheme differs in that the residual that is used comes from the system just solved and not from the original one; that is, $r_0=b-Az$ and if d_0 is the solution of $Ax=r_0$, then $r_1=r_0-Ad_0$. In general we have

$$r_{k+1}=r_k-Ad_k$$

and d_k is the approximate solution of $Ax=r_k$.

THEOREM 4.1. *Let the vectors* r_k *and* d_k *be defined as above. Then if* $N(r_{k+1}) < aN(r_k), 0 < a < 1$, r_k *approaches zero as* k *goes to infinity and* $z+d_0+d_1+\ldots+d_k$ *converges to* $A^{-1}b$.

PROOF: We have

$$r_0 = b - Az$$

$$r_1 = r_0 - Ad_0$$

$$r_2 = r_1 - Ad_1$$

$$\cdot \quad \cdot \quad \cdot$$

$$r_{k+1} = r_k - Ad_k$$

so that

$$r_0 + r_1 + \cdots + r_{k+1} = b + r_0 + r_1 + \cdots + r_k - A(z + d_0 + d_1 + \cdots + d_k)$$

or

$$r_{k+1} = b - A(z + d_0 + d_1 + \cdots + d_k).$$

Now

$$N(r_{k+1}) < aN(r_k) < a^2N(r_{k-1}) < \cdots < a^{k+1}N(r_0).$$

Hence, as k increases to infinity, a^{k+1} goes to zero, and then $N(r_{k+1})$ goes to zero which implies that r_{k+1} goes to zero. Therefore $A(z + d_0 + d_1 + \cdots + d_k)$ approaches b which means that $z + d_0 + d_1 + \cdots + d_k$ converges to $A^{-1}b$.

Instead of summing the residuals, if we use Wilkinson's backward approach, we could say that d_k is the exact solution of a slightly perturbed system and we would have

$$(A + E_k)d_k = r_k$$

where we have subscripted the E to show the dependence on each system. With the introduction of E_k we have another convergence criterion.

THEOREM 4.2. *Let the vectors* r_k, d_k *and the matrix* E_k *be defined as above. Assume* $N(E_k) < c$ *for all* k. *Then if* $N(d_{k+1}) < pN(d_k)$, $0 < p < 1$, r_k *approaches zero as* k *goes to infinity and* $z + d_0 + d_1 + \cdots + d_k$ *converges to* $A^{-1}b$.

PROOF:

$$(A + E_k)d_k = r_k$$

$$Ad_k + E_kd_k = r_k$$

$$E_kd_k = r_k - Ad_k$$

but

$$r_{k+1} = r_k - Ad_k$$

so that

$$r_{k+1} = E_kd_k$$

and

$$N(r_{k+1}) \leq N(E_k)N(d_k).$$

Now

$$N(d_k) < pN(d_{k-1}) < p^2N(d_{k-2}) < \cdots < p^kN(d_0).$$

Hence,

$$N(r_{k+1}) < cp^kN(d_0)$$

so that as k goes to infinity, p^k approaches zero and r_{k+1} approaches zero and the conclusion follows just as in Theorem 4.1.

COROLLARY. *Under the same conditions as the theorem, we have* $N[A^{-1}b-(z+d_0+\ldots+d_k)] \leq \frac{p^{k+1}}{1-p} N(d_0)$.

PROOF:

$$A^{-1}b-(z+d_0+\ldots+d_k)=d_{k+1}+d_{k+2}+\ldots$$

$$N[A^{-1}b-(z+d_0+\ldots+d_k)] \leq N(d_{k+1})+N(d_{k+2})+\ldots$$

$$\leq p^{k+1}N(d_0)+p^{k+2}N(d_0)+\ldots$$

$$=p^{k+1}(1+p+p^2+\ldots)N(d_0)$$

$$=\frac{p^{k+1}}{1-p}N(d_0).$$

We will see that these two conditions are actually contained in our next theorem. There $N(I-XA)<1$ is our condition, and this is the p here; and also $N(I-AX)<1$ is used, and this is the a here.

It would seem, however, that repeatedly calculating the residuals from the primary equation, $b-Ax$, is a better scheme since the magnitude of the residuals that are being used may be much smaller for the alternate method. Using the smaller numbers might demand multiprecision routines, more than the double precision necessary for the calculation of each residual in the standard approach. On the other hand, it has been found that $N(r_{k+1})$ is not always smaller than $N(r_k)$ for the standard scheme, and in similar problems using the other scheme we have found that the residuals do decrease at each step. This could be due to the fact that in the second scheme we make one addition at the end, $z+d_0+d_1+\ldots+d_k$, for our final answer and thus only one rounding error occurs. In the standard scheme, however, this is done at each step and if this rounding error is included in the convergence proof, we would find that the error, $A^{-1}b-x_k$, does not approach zero but only gets small [12, p. 110]. Hence, this convergence proof is really for what Forsythe calls an "Almost Theorem" [13, p. 9], especially since Kahan has found a counterexample to the standard approach [10, p. 78]. In practice we have found both schemes produce identical results.

Let us now turn our attention to a different convergence proof for the standard scheme. It will be useful to write out a few steps of the process. Let d_k be the computed solution at each step.

(1) solve $Ax=b$:

$$z=Xb$$

$$r=b-Az$$

$$=b-A(Xb)$$

$$=(I-AX)b$$

(2) solve $Ax=r$:

$$d_0=Xr$$

$$=X(b-Az)$$

$$=Xb-XAz$$

$$=(I-XA)z$$

$$=Yz,\ Y=I-XA.$$

improved solution:

$$\begin{aligned} z_1 &= z + d_0 \\ &= z + Yz \\ &= (I+Y)z \end{aligned}$$

and

$$\begin{aligned} r_1 &= b - Az_1 \\ &= b - A(z+d_0) \\ &= b - Az - Ad_0 \\ &= r - AXr \\ &= (I-AX)r. \end{aligned}$$

(3) solve $Ax = r_1$:

$$\begin{aligned} d_1 &= Xr_1 \\ &= X(I-AX)r \\ &= (I-XA)Xr \\ &= YXr \\ &= Y^2z \end{aligned}$$

improved solution:

$$\begin{aligned} z_2 &= z_1 + d_1 \\ &= (I+Y)z + Y^2z \\ &= (I+Y+Y^2)z \end{aligned}$$

and

$$\begin{aligned} r_2 &= b - Az_2 \\ &= b - A(z_1+d_1) \\ &= b - Az_1 - Ad_1 \\ &= r_1 - AXr_1 \\ &= (I-AX)r_1 \\ &= (I-AX)^2r. \end{aligned}$$

In general, we have

$$\begin{aligned} d_{k-1} &= Xr_{k-1} \\ &= Yd_{k-2} \\ &= Y^kz, \end{aligned}$$

$$\begin{aligned} z_k &= z_{k-1} + d_{k-1} \\ &= (I+Y+Y^2+\ldots+Y^k)z, \end{aligned}$$

$$\begin{aligned} r_k &= (I-AX)r_{k-1} \\ &= (I-AX)^kr \end{aligned}$$

406-660 O - 71 - 4

for by induction

$$\begin{aligned} d_k &= Xr_k \\ &= X(I-AX)^k r \\ &= X(I-AX)(I-AX)^{k-1} r \\ &= YXr_{k-1} \\ &= Yd_{k-1} \\ &= Y^{k+1} z; \end{aligned}$$

$$\begin{aligned} z_{k+1} &= z_k + d_k \\ &= (I+Y+Y^2+\ldots+Y^k)z + Y^{k+1}z \\ &= (I+Y+Y^2+\ldots+Y^k+Y^{k+1})z; \end{aligned}$$

$$\begin{aligned} r_{k+1} &= b - Az_{k+1} \\ &= b - A(z_k + d_k) \\ &= b - Az_k - Ad_k \\ &= r_k - AXr_k \\ &= (I-AX)r_k \\ &= (I-AX)^{k+1} r. \end{aligned}$$

THEOREM 4.3. *Let the vectors* d_k, z_k, r_k *and the matrix* Y *be defined as above. Then if* $N(Y) < 1$, z_k *converges to* $A^{-1}b$ *as* k *goes to infinity.*

PROOF: $z_k = (I+Y+Y^2+\ldots+Y^k)z$ and, as we proved in Lemma 1.2, if $N(Y) < 1$, $I+Y+Y^2+\ldots+Y^k$ converges to $(I-Y)^{-1}$ as k goes to infinity. Also

$$Y = I - XA$$

$$XA = I - Y$$

$$A^{-1}X^{-1} = (I-Y)^{-1}.$$

Hence, as k goes to infinity, z_k converges to

$$(I-Y)^{-1}z = (A^{-1}X^{-1})(Xb) = A^{-1}b.$$

COROLLARY 1. *Let* d_k *and* r_k *be defined as above. Then if* $N(Y) < 1$, d_k *goes to zero as* k *goes to infinity and if* $N(I-AX) < 1$, r_k *also goes to zero.*

PROOF:

$$d_{k-1} = Y^k z$$

so that

$$N(d_{k-1}) \leq N(Y)^k N(z).$$

Since $N(Y) < 1$, the righthand side goes to zero as k goes to infinity so that the lefthand side does also, which implies that d_{k-1} goes to zero. Since

$$r_k = (I - AX)^k r$$

and

$$N(r_k) \leq N(I - AX)^k N(r),$$

if $N(I - AX) < 1$, the same argument implies that r_k goes to zero as k goes to infinity.

We still have linear convergence but we feel using $N(Y) < 1$ is more meaningful than $N(A^{-1}E) < \frac{1}{2}$. Since we need X if we are going to use any bounds for $N(A^{-1}b - z_k)$, it is also a practical criterion.

From this theorem we can also get an estimate of the improvement at each step as well as a final error bound.

COROLLARY 2. *With the same definitions and conditions as in the theorem,*

$$(a) \quad \mathrm{N}(z_k - z_{k-1}) \leq \mathrm{N}(Y)^k \mathrm{N}(z) \text{ and}$$

$$(b) \quad \mathrm{N}(A^{-1}b - z_k) \leq \frac{\mathrm{N}(Y)^{k+1}}{1 - \mathrm{N}(Y)} \mathrm{N}(z)$$

PROOF:

(a) $$z_k - z_{k-1} = d_{k-1}$$
$$= Y^k z$$

so that

$$N(z_k - z_{k-1}) \leq N(Y)^k N(z).$$

(b) $$A^{-1}b - z_k = (I - Y)^{-1} z - z_k$$
$$= (I - Y)^{-1} z - (I + Y + \ldots + Y^k) z$$
$$= (Y^{k+1} + Y^{k+2} + \ldots) z$$

so that

$$N(A^{-1}b - z_k) \leq \frac{N(Y^{k+1})}{1 - N(Y)} N(z)$$
$$\leq \frac{N(Y)^{k+1}}{1 - N(Y)} N(z).$$

This is a smaller number generally than that bounding the difference between two successive steps in the iteration. In fact, if $N(Y) \leq \frac{1}{2}$,

$$\frac{N(Y)}{1 - N(Y)} \leq 1$$

and

$$N(A^{-1}b - z_k) \leq \frac{N(Y)^{k+1}}{1 - N(Y)} N(z)$$
$$= N(Y)^k N(z) \cdot \frac{N(Y)}{1 - N(Y)}$$
$$\leq N(Y)^k N(z).$$

We note that for the convergence of the sequence of z_k we need only that the norm of the lefthand residual be less than one. This implies that the residuals, $b - Az_k$, need not go to zero as k

increases and this has been the case in some problems as previously mentioned. This fact also emphasizes that in the solution of $Ax=b$, the lefthand inverse is the more important one. As was pointed out before, a good right inverse may not imply that the error in the solution vector is small. Hence, if we intend to use an approximate inverse in our solution of $Ax=b$ it is very important that the improvement scheme used be one that improves the lefthand inverse. This is very easily done by the methods of the previous section.

5. Error Bounds for the Approximate Inverse of Special Matrices

In this section we will briefly discuss some results concerning a symmetric matrix, derive an error bound for the inverse of a triangular matrix, and then investigate the LU decomposition of a matrix A for possible error bounds on A^{-1} which we will apply to the solution of $Ax=b$.

Since we will be developing results that apply to some special matrices, it might be good to give some properties of these matrices so that they will be readily available. We will not give any proofs as these can be found in various texts on linear algebra or numerical linear algebra as listed in the bibliography.

Let A be a real symmetric matrix.

1. The eigenvalues of A^2 are the squares of the eigenvalues of A.

2. The eigenvalue of maximum modulus of A^*A is the square of that of A. (A^* is the conjugate transpose of A and hence, in our case, is just the transpose of A, A^T.)

3. The spectral norm of any matrix A, $N_s(A)$, is defined as the square root of the eigenvalue of maximum modulus of A^*A. Hence, in our case, $N_s(A)$ is just the absolute value of the eigenvalue of A which has maximum modulus.

4. A is orthogonally similar to a diagonal matrix; i.e., $A=O^TDO$ where O is orthogonal, $O^T=O^{-1}$ and D is diagonal.

We will also use a few general relations which we will mention at this time.

1. Let $N_f(A)$ be the Frobenius norm of A. Then if A and B are orthogonally similar, $N_f(A)=N_f(B)$. (See section 1, where we proved this for unitarily similar matrices.)

2. We recall a theorem of Schur (also used in sect. 1): If A is an arbitrary matrix, then it can be transformed to a triangular matrix by means of a unitary matrix; i.e., a unitary matrix U exists such that $U^*AU=T$, where $U^*=U^{-1}$ and T is upper triangular.

We will now define a relation between matrices which may not be too familiar.

DEFINITION. A *is dominated by* B *if and only if* $|a_{ij}| \leq b_{ij}$ *for all* i *and* j. *Symbolically, we write* $A \ll B$. *Some basic properties of this relation are that if* $A_1 \ll B_1$ *and* $A_2 \ll B_2$, *then* $A_1+A_2 \ll B_1+B_2$ *and* $A_1A_2 \ll B_1B_2$.

Finally, we will use the symbol J to represent the $n \times n$ matrix whose elements are all ones, and l_i will be used for the eigenvalues of the matrix under consideration.

With this background we can proceed to our first theorem.

THEOREM 5.1. *If* A *is a real symmetric matrix and is* $0 < r \leq |l_1|$, *then* $N_s(A^{-1}) \leq 1/r$.

PROOF:

$$r \leq |l_i|$$

$$\frac{1}{|l_i|} \leq \frac{1}{r}$$

$$|l_i(A^{-1})| \leq 1/r$$

$$\max |l_i(A^{-1})| \leq 1/r$$

and

$$N_s(A^{-1}) \leq 1/r.$$

In our next few theorems let the following notation hold. Let X be the approximate inverse of A and let $R=OXO^T=(r_{ij})$ where O is an orthogonal matrix such that $A=O^TDO$.

THEOREM 5.2. *Let* A *be a real symmetric matrix. Then if*

$$\mathrm{N}_f(\mathrm{I}-\mathrm{AX}) \leqslant \mathrm{e}, \quad \mathrm{N}_f(\mathrm{A}^{-1}-\mathrm{X}) \leqslant \mathrm{n}^{1/2}\frac{\mathrm{e}}{\min|\mathrm{l}_i|}.$$

PROOF:

$$A^{-1}-X = O^T D^{-1} O - O^T R O$$
$$= O^T(D^{-1}-R)O$$

and

$$N_f(A^{-1}-X) = N(D^{-1}-R)$$
$$= N[D^{-1}(I-DR)].$$

Now
$$I-AX = I-O^T DRO$$
$$= O^T(I-DR)O$$

so that
$$N_f(I-AX) = N_f(I-DR).$$

Hence
$$N_f(A^{-1}-X) \leqslant N_f(D^{-1})N_f(I-AX).$$

Now
$$D^{-1} \leqslant \frac{1}{\min|l_i|} I$$

so that
$$N_f(D^{-1}) \leqslant \frac{n^{1/2}}{\min|l_i|}$$

and if $N_f(I-AX) \leqslant e$, we have

$$N_f(A^{-1}-X) \leqslant n^{1/2}\frac{e}{\min|l_i|}.$$

COROLLARY 1. *Under the same conditions as the theorem except that if* $\mathrm{I}-\mathrm{AX} \leqslant \mathrm{kJ}$, *we have*

$$\mathrm{N}_f(\mathrm{A}^{-1}-\mathrm{X}) \leqslant \mathrm{n}^{3/2}\frac{\mathrm{k}}{\min|\mathrm{l}_i|}.$$

PROOF: Since $N_f(J)=n$, we use nk instead of e as in the theorem.

COROLLARY 2. *Let all the eigenvalues of* A *be the same, say a. Then*

$$N_f(A^{-1}-X) \leqslant \frac{e}{|a|}.$$

PROOF: We use the fact that $D=aI$ and the property of a norm.

THEOREM 5.3. *Let the* n *eigenvalues of* A *be distinct. Let* $\mathrm{E}=(\mathrm{I}-\mathrm{XA})-(\mathrm{I}-\mathrm{AX})$ *and let* $\mathrm{E}_1=\mathrm{OEO}^{\mathrm{T}}=(\mathrm{e}_{ij})$. *Then if* $\mathrm{d}_i=\mathrm{l}_i^{-1}-\mathrm{r}_{ii}$,

$$\mathrm{N}_f(\mathrm{A}^{-1}-\mathrm{X}) \leqslant \mathrm{n}\ \max\left(|\mathrm{d}_i|, \left|\frac{\mathrm{e}_{ij}}{\mathrm{l}_i-\mathrm{l}_j}\right|\right).$$

PROOF:

$$E=(I-XA)-(I-AX)$$
$$=AX-XA$$
$$=O^T(DR-RD)O.$$

If

$$E_1=OEO^T, \qquad E_1=DR-RD.$$

In terms of elements, this last equation can be written

$$e_{ij}=l_ir_{ij}-l_jr_{ij}, \quad i\neq j$$
$$=0 \qquad\quad , \quad i=j$$

or

$$r_{ij}=\frac{e_{ij}}{l_i-l_j} \quad , \quad i\neq j$$

and for $i=j$, r_{ij} is not determined but they should be close to l_i^{-1}. Now consider $D^{-1}-R$. The diagonal elements are d_i and the off-diagonal, $-e_{ij}/(l_i-l_j)$. Then

$$D^{-1}-R \ll \max\left(|d_i|, \left|\frac{e_{ij}}{l_i-l_j}\right|\right)J$$

so that

$$N_f(A^{-1}-X)=N_f(D^{-1}-R)$$
$$\leq n \max\left(|d_i|, \left|\frac{e_{ij}}{l_i-l_j}\right|\right).$$

The spread of the eigenvalues of a matrix can give a good idea of the condition of a matrix; in fact, the ratio of the modulus of the largest to that of the smallest is defined as the P-condition of a matrix. We see from our error bound that a large spread does not necessarily mean we cannot get a good approximate inverse. This bound implies, however, that if two eigenvalues are very close, we might expect trouble but that this might be counteracted if E is small enough. Hence, once again knowledge of the lefthand residual matrix is useful.

COROLLARY. *IF* $\mathrm{E}=0$, $\mathrm{N}_f(\mathrm{A}^{-1}-\mathrm{X})\leq \mathrm{n}\max_i|\mathrm{d}_i|$.

PROOF: If $E=0$, $E_1=0$ and then $r_{ij}=0$ for $i\neq j$. Hence

$$D^{-1}-R=\text{diagonal } (d_i)$$
$$\ll \max |d_i| J$$

so that

$$N_f(A^{-1}-X)\leq n \max |d_i|.$$

Let us now consider another special case, a triangular matrix T, and derive an error bound for its inverse. Let N_{rs} be the maximum row sum norm.

THEOREM 5.4. *Let* T *be a triangular matrix and let* $\mathrm{e}=\max_i \left|\frac{\mathrm{t_{ij}}}{\mathrm{t_{ii}}}\right|$ *for all* j. *Then*

$$\mathrm{N_{rs}}(\mathrm{T}^{-1})\leq \frac{(1+\mathrm{e})^{n-1}}{\min|\mathrm{t_{ii}}|}.$$

PROOF: Let $T=D-U=D(I-D^{-1}U)=D(I-E)$, D diagonal, U strictly upper and $E=D^{-1}U=(e_{ij})$ where $e_{ij}=t_{ij}/t_{ii}$. Then $T^{-1}=(I-E)^{-1}D^{-1}$. Let us examine $(I-E)^{-1}$. Since E has zero diagonal, $E^n=0$ and $(I-E)^{-1}=I+E+E^2+\ldots+E^{n-1}$. Let $e=\max_i |e_{ij}|$, for all j. Then we can write $E \ll eP$ where

$$P=\begin{bmatrix} 0 & 1 & 1 & 1 & \ldots & 1 \\ & 0 & 1 & 1 & \ldots & 1 \\ & & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \vdots \\ & & & & \ddots & \vdots \\ & & & & \ddots & \vdots \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix}.$$

Let us examine the various powers of P.

$$P^2=\begin{bmatrix} 0 & 0 & 1 & 2 & 3 & 4 & \ldots & n-2 \\ & 0 & 0 & 1 & 2 & 3 & \ldots & n-3 \\ & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & 2 \\ & & & & & & \ddots & 1 \\ & & & & & & & 0 \\ & & & & & & & 0 \end{bmatrix}$$

$$P^3=\begin{bmatrix} 0 & 0 & 0 & 1 & 3 & 6 & 10 & 15 & \ldots & \binom{n-2}{2} \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & 10 \\ & & & & & \ddots & \ddots & \ddots & \ddots & 6 \\ & & & & & & \ddots & \ddots & \ddots & 3 \\ & & & & & & & \ddots & \ddots & 1 \\ & & & & & & & & \ddots & 0 \\ & & & & & & & & & 0 \\ & & & & & & & & & 0 \end{bmatrix}$$

In P^3 we have written $\binom{n-2}{2}$ instead of $1+2+3+\ldots+n-3$ which comes from multiplying the first row of P times the nth column of P^2. We also note that $n-2$ can be written $\binom{n-2}{1}$ which we will use in P^2; of course $\binom{n-2}{0}=1$, the element in the $(1, n)$ position of P^1. Generalizing this pattern and realizing that $P^n=0$, we find that P^{n-1} has all zeros except for a 1 in the $(1, n)$ position which we will write as $\binom{n-2}{n-2}$. Let us write the first row (only) of the next few powers:

$$P_1^4=\left[0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 4 \quad 10 \quad 20 \quad \ldots \quad \binom{n-2}{3}\right]$$

$$P_1^5=\left[0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 5 \quad 15 \quad \ldots \quad \binom{n-2}{4}\right].$$

We restrict ourselves to the maximum row sum norm and our interest is in

$$N(I+E+E^2+\ldots+E^{n-1}).$$

As mentioned in our first chapter, it will be better to add the matrices first and then take the norm. It is clear that the maximum row sum will come from the first row. Let us rewrite the first rows of these matrices, then, using the binomial expression $\binom{n}{k}$.

$$P_1^1=\left[0 \quad \binom{0}{0} \quad \binom{1}{0} \quad \binom{2}{0} \quad \binom{3}{0} \quad \binom{4}{0} \quad \binom{5}{0} \quad \ldots \quad \binom{n-2}{0}\right]$$

$$P_1^2=\left[0 \quad 0 \quad \binom{1}{1} \quad \binom{2}{1} \quad \binom{3}{1} \quad \binom{4}{1} \quad \binom{5}{1} \quad \ldots \quad \binom{n-2}{1}\right]$$

$$P_1^3=\left[0 \quad 0 \quad 0 \quad \binom{2}{2} \quad \binom{3}{2} \quad \binom{4}{2} \quad \binom{5}{2} \quad \ldots \quad \binom{n-2}{2}\right]$$

$$P_1^4=\left[0 \quad 0 \quad 0 \quad 0 \quad \binom{3}{3} \quad \binom{4}{3} \quad \binom{5}{3} \quad \ldots \quad \binom{n-2}{3}\right]$$

We have $E \ll eP$ so that $E^k \ll e^k P^k$. Let us take the sum of the various powers position by position; i.e., find the sum for the (1, 1) position, (1, 2) position, . . ., $(1, n)$ position.

$(1, 1)$: $1 = 1$

$(1, 2)$: $\binom{0}{0} e = e(1+e)^0$

$(1, 3)$: $\binom{1}{0} e + \binom{1}{1} e^2 = e(1+e)^1$

$(1, 4)$: $\binom{2}{0} e + \binom{2}{1} e^2 + \binom{2}{2} e^3 = e(1+e)^2$

$(1, 5)$: $\binom{3}{0} e + \binom{3}{1} e^2 + \binom{3}{2} e^3 + \binom{3}{3} e^4 = e(1+e)^3$

. . .

$(1, n)$: $\binom{n-2}{0} e + \binom{n-2}{1} e^2 + \binom{n-2}{2} e^3 + \ldots + \binom{n-2}{n-2} e^{n-1} = e(1+e)^{n-2}$

The norm, therefore, comes from adding the extreme righthand column:

$$\begin{aligned} N_{rs}[(I-E)^{-1}] &= 1 + e(1+e)^0 + e(1+e)^1 + e(1+e)^2 + \cdots + e(1+e)^{n-2} \\ &= 1 + e[1 + (1+e) + (1+e)^2 + \cdots + (1+e)^{n-2}] \\ &= 1 + e\left[\frac{(1+e)^{n-1} - 1}{e}\right] \\ &= (1+e)^{n-1}. \end{aligned}$$

Hence we have

$$\begin{aligned} N_{rs}(T^{-1}) &\leq N_{rs}(D^{-1})(1+e)^{n-1} \\ &= \frac{(1+e)^{n-1}}{\min\limits_i |t_{ii}|} \end{aligned}$$

and all the quantities are readily available by inspection of T.

COROLLARY 1.

$$N_{rs}(T^{-1}) \leq \frac{\left(\min\limits_i |t_{ii}| + \max\limits_i |t_{ij}|\right)^{n-1}}{\min\limits_i |t_{ii}|^n}.$$

PROOF: We use the fact that

$$\max_i \left|\frac{t_{ij}}{t_{ii}}\right| \leq \frac{\max\limits_i |t_{ij}|}{\min\limits_i |t_{ii}|} \quad \text{for all } j$$

and substitute this inequality in the bound of the theorem.

We see in this last bound that the two important numbers are the smallest and the largest. This ties in very well with the concept of the condition of a matrix, because a matrix with a large spread in the magnitude of its elements is generally considered an ill-conditioned matrix. The larger the difference of the maximum element and the minimum element, the higher the value of this bound.

COROLLARY 2.

$$\textit{If } e \leq 1, \quad N_{rs}(T^{-1}) \leq \frac{2^{n-1}}{\min |t_{ii}|}.$$

This corollary would definitely be applicable to a diagonally dominant triangular matrix.

THEOREM 5.5. *Let* A *be any matrix and let* e *be defined as in Theorem 5.4. Then if* $N_{rs}(I - AX) \leq q$,

$$N_f(A^{-1} - X) \leq nq \frac{(1+e)^{n-1}}{\min |l_i|}.$$

PROOF: If U is a unitary matrix and T triangular, we have

$$\begin{aligned} A &= U^*TU \\ X &= U^*RU, \qquad \text{for some } R \\ A^{-1} - X &= U^*(T^{-1} - R)U \end{aligned}$$

so that

$$\begin{aligned} N_f(A^{-1}-X) &= N_f(T^{-1}-R) \\ &= N_f(T^{-1}(I-TR)) \\ &\leq N_f(T^{-1})N_f(I-TR) \\ &= N_f(T^{-1})N_f(I-AX) \end{aligned}$$

following the same arguments that were used in the proof of Theorem 5.2.

Now for any A we have the relation $N_f(A) \leq n^{1/2}N_{rs}(A)$. Using this inequality we continue

$$N_f(A^{-1}-X) \leq nN_{rs}(T^{-1})N_{rs}(I-AX).$$

Using our bounds for $N_{rs}(T^{-1})$ we have

$$N_f(A^{-1}-X) \leq n\frac{(1+e)^{n-1}}{\min|l_i|}N_{rs}(I-AX)$$

where $e=\max_i\left|\frac{t_{ij}}{t_{ii}}\right|$ for all j, and l_i are the eigenvalues of A. The difficulty here, of course, is our lack of knowledge of t_{ij} but the bound certainly brings out the fact that the smaller the eigenvalue of A, the higher the bound. This also implies that for a nearly singular matrix (a zero eigenvalue), the bounds are much higher which gives another indication of the influence of ill conditioning on the calculation of the inverse.

A more practical application of the bounds for the inverse of a triangular matrix might be made to the error in the solution, z, of a linear algebraic system, $Ax=b$. We pointed out in section 2 that to get a bound on the error, $A^{-1}b-z$, we must use an approximate inverse. This necessitates solving an additional n systems. However, if we use the LU decomposition method (Gaussian Elimination), we need not do all that. In solving our one given system, the first step is to decompose A into the product of an upper triangular matrix U, and a lower triangular matrix L, whose diagonal elements are all ones.

$$A=LU$$

Then

$$A^{-1}=U^{-1}L^{-1}$$

and

$$N(A^{-1}) \leq N(U^{-1})N(L^{-1}).$$

Using the bounds just derived for a triangular matrix, we have

$$N(U^{-1}) \leq \frac{(1+e_1)^{n-1}}{\min|u_{ii}|}, \quad e_1=\max_i\left|\frac{u_{ij}}{u_{ii}}\right| \text{ for all } j$$

and

$$N(L^{-1}) \leq \frac{(1+e_2)^{n-1}}{1}, \quad e_2=\max_{ij}|l_{ij}|.$$

The quantities in the last wo bounds are readily available from the decomposition of A and thus with very little extra work we will have a bound for the error in the solution vector z without needing an approximate inverse. Since for any norm, $N(A^{-1}b-z) \leq N(A^{-1})N(b-Az)$, we have

THEOREM 5.6.

$$N_{rs}(A^{-1}b-z) \leq \frac{[(1+e_1)(1+e_2)]^{n-1}}{\min|u_{ii}|}N_{rs}(b-Az).$$

How much of an overestimate this may be again results primarily from using

$$N(AB) \leq N(A)N(B)$$

as mentioned in section 1. We know that the residual is usually much smaller than the actual error, however, and this fact may help make this bound practical.

Examining in itself the LU decomposition of A will give us a slight refinement in the bound for L^{-1}. In the decomposition process we multiply A on the left by a succession of elementary row operators to reduce it to triangular form:

$$M_{n-1}M_{n-2} \cdots M_1 A = U.$$

Hence, $U^{-1}M_{n-1}M_{n-2} \cdots M_1 A = U^{-1}U = I$ and therefore

$$A^{-1} = U^{-1}M_{n-1}M_{n-2} \cdots M_1$$

so that

$$\begin{aligned} L^{-1} &= M_{n-1}M_{n-2} \cdots M_1 \\ &= M \end{aligned}$$

and

$$L = M_1^{-1}M_2^{-1} \cdots M_{n-1}^{-1}.$$

Now each M_k is a lower triangular matrix with elements in only the kth column and ones along the diagonal:

$$M_k = \begin{bmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & & & & \\ & & & -m_{k+1,k} & 1 & & & \\ & & & -m_{k+2,k} & & \ddots & & \\ & & & \vdots & & & \ddots & \\ & & & -m_{nk} & & & & 1 \end{bmatrix}$$

so M_k^{-1} is actually M_k with the signs of the off-diagonal elements reversed. Furthermore

$$L = \begin{bmatrix} 1 & & & & & \\ m_{21} & 1 & & & & \\ m_{31} & m_{32} & 1 & & & \\ m_{41} & m_{42} & \ddots & \ddots & & \\ \vdots & \vdots & & & \ddots & \\ m_{n1} & m_{n2} & & & m_{n,n-1} & 1 \end{bmatrix}.$$

Now $N(M_k) \leq 1 + \max_i |m_{ik}| \leq 1 + \max_{ij} |m_{ij}|$ for each k so that

$$N(L^{-1}) = N(M)$$
$$\leq \left(1 + \max_{ij} |m_{ij}|\right)^{n-1}$$

and this is exactly what we have found from our other approach for a triangular matrix. (Note that $l_{ij} = m_{ij}$ and these are stored in the memory of the computer.) This could be a large overestimate since we are repeatedly using the inequality we have indicated as a source of trouble:

$$N(M) \leq N(M_{n-1})N(M_{n-2}M_{n-3} \ldots M_1) \leq \ldots \leq N(M_{n-1})N(M_{n-2}) \ldots N(M_1).$$

The refinement comes from not using the second inequality indicated for $N(M_k)$. Let $p_j = \max_i |m_{ij}|$ for each j. Then

$$N(M_k) = 1 + p_k$$

and

$$N(L^{-1}) = N(M) \leq \prod_{j=1}^{n-1} (1 + p_j).$$

Hence we have

$$N(A^{-1}) \leq N(L^{-1})N(U^{-1})$$
$$\leq \prod_{j=1}^{n-1} (1 + p_j) \frac{(1 + e_2)^{n-1}}{\min_i |u_{ii}|},$$

where $e_2 = \max_i \left|\frac{u_{ij}}{u_{ii}}\right|$ for all j.

Finally,

$$N(A^{-1}b - z) \leq \prod_{j=1}^{n-1} (1 + p_j) \frac{(1 + e_2)^{n-1}}{\min |u_{ii}|} N(b - Az)$$

where the notation is the same as before and N is the maximum row sum norm. We therefore have a bound on the error of the approximate solution of a linear algebraic system without the knowledge of the inverse of the system.

Let us now give a numerical example of this bound and also of those from section 2. We will use the example giver in J. H. Wilkinson's book *Rounding Errors in Algebraic Processes*, page 111, and we refer the reader to the discussion given there concerning the size of the bounds for the two different values of the righthand sides.

Problem 1:

$$N(A) = .252E+01 \qquad N(b_1) = .912E+00 \qquad N(A^{-1}b_1) = .225E+05$$
$$N(A^{-1}) = .150E+06 \qquad N(b_1 - Az) = .711E-01 \qquad N(A^{-1}b_1 - z) = .151E+04$$

$$1.\ N(A^{-1}) \leq \prod_{j=1}^{n} (1 + p_j) \frac{(1 + e_2)^{n-1}}{\min |u_{ii}|} = P = .867E+06$$
$$\leq \frac{[(1 + e_1)(1 + e_2)]^{n-1}}{\min |u_{ii}|} = .112E+07$$
$$N(X) = .161E+06$$

2. $N(A^{-1}b_1-z) \leq N(A^{-1})N(b_1-Az) \quad = D = .107E+05$

$$\text{or } PN(b_1-Az) \quad = E = .616E+05$$

$$\text{or } \frac{N(X)N(b_1-Az)}{1-N(Y)} \quad = F = .138E+05$$

$$\text{or } \frac{N(X)N(Y)N(b_1)}{1-N(Y)} \quad = G = .307E+05$$

3. $\dfrac{N(A^{-1}b_1-z)}{N(A^{-1}b_1)} \leq \dfrac{D}{N(A^{-1}b_1)} \quad = .475E+00$

$$\text{or } \frac{E}{N(A^{-1}b_1)} \quad = .274E+01$$

$$\text{or } \frac{F}{N(A^{-1}b_1)} \quad = .614E+00$$

$$\text{or } \frac{G}{N(A^{-1}b_1)} \quad = .137E+01$$

$$\text{or } N(I-XA) \quad = .125E+00$$

$$= \text{actually} \quad = .673E-01$$

4. $N(A^{-1}b_1-z) \geqslant \dfrac{N(b_1-Az)}{N(A)} \quad = .283E-01$

Problem 2:

$N(A) = .252E+01 \qquad N(b_2) = .876E+00 \qquad N(A^{-1}b_2) = .915E+00$

$N(A^{-1}) = .150E+06 \qquad N(b_2-Az) = .231E-06 \qquad N(A^{-1}b_2-z) = .607E-02$

(Same as previous problem except that the third element of b_1 was decreased by .264139.)

1. $N(A^{-1}b_2-z) \leq N(A^{-1})N(b_2-Az) \quad = D = .348E-01$

$$\text{or } PN(b_2-Az) \quad = E = .200E+00$$

$$\text{or } \frac{N(X)N(b_2-Az)}{1-N(Y)} \quad = F = .449E-01$$

$$\text{or } \frac{N(X)N(Y)N(b_2)}{1-N(Y)} \quad = G = .295E+05$$

$$\geqslant \frac{N(b_2-Az)}{N(A)} \quad = H = .919E-07$$

2. $\dfrac{N(A^{-1}b_2-z)}{N(A^{-1}b_2)} \leq \dfrac{D}{N(A^{-1}b_2)} \quad = .380E-01$

$$\text{or } \frac{E}{N(A^{-1}b_2)} \quad = .219E+00$$

$$\text{or } \frac{F}{N(A^{-1}b_2)} \quad = .491E-01$$

$$\text{or } \frac{G}{N(A^{-1}b_2)} \quad = .322E+05$$

$$\text{or } N(I-XA) \quad = .125E+00$$

$$= \text{actually} \quad = .668E-02$$

6. Evaluation of Computer Programs

In this section, the second part of our research, we would like to use the mathematics we have developed in our first five sections in the evaluation of the performance of some computer programs. We shall set up a problem, choose some programs, and select some test cases to use as the data for each program.

Since calculating the inverse of a matrix usually includes the solution of a linear algebraic system, we have chosen the inversion of a matrix for our problem. The programs we will use were chosen primarily on the basis of availability. Some were on the computer library tapes at the University of Maryland, some were in the literature, and others were received from individuals in the field. It seems that almost all of the programs being used today employ some form of Gauss elimination to do matrix inversion, so we have confined ourselves to this one direct method. The test matrices were chosen to cover a variety of condition numbers, from a small value where all programs should perform quite well, to higher values where some difficulties could be encountered.

Our main concern will be with the accuracy of the result; i.e., with the smallness of $N(A^{-1}-X)$, where X is the approximate inverse produced by a computer program. (This, of course, is not the only criterion of success but only the originator of the problem can decide exactly what is important to him.) Hence, we use the error bound derived in section 1, that

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)}, \qquad Y = I - AX,$$

as our basis for evaluation. We will also limit ourselves to the righthand residual matrix since it is usually smaller, although, as will be seen, this is not always true. It has been frequently pointed out that theoretical error estimates are generally overestimates and sometimes of little practical value. For comparison purposes, however, they serve quite well and furthermore, we will show that this particular error bound is actually quite close to reality. It is true that we have considered only a sampling of computer programs and test problems but we feel they are quite representative and that further work will only support our conclusions.

There are many ways of arriving at a "condition number" for a matrix. The one we shall use is called the P-condition number:

$$P(A) = \left|\frac{\lambda}{\mu}\right|$$

where λ is an eigenvalue of largest modulus and μ, of smallest [14].

Before going into the details of the test matrices and the computer programs, there are some very important items that should be noted first. As we have indicated in many of the previous sections, the possibility of the occurrence of some kind of error must always be kept in mind. Checking the input, execution, and the output of a computer run may be tedious but it is quite essential. We have chosen matrices that can be generated within the machine, used two kinds of checks throughout the execution, and have the exact inverses available to check the final results. The main check on the execution of the program (and the machine) is achieved by including an extra row and column in the original matrix before entering the inversion routine. If the additional column is made up of the negative of the sum of the elements of each of the rows of the matrix, then the row sums of the augmented matrix should always be zero. If we put a 1 in the $(n+1, n+1)$ position and zeros in the other positions of the $n+1$ row, the last column in the computed inverse should be all ones. Let $b = (1, 1, \ldots, 1)^T$. Then

$$\begin{bmatrix} A & -Ab \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & b \\ 0 & 1 \end{bmatrix}.$$

Hence, the maximum difference from unity will give us a good indication if anything has gone wrong. What may have happened (machine failure, large rounding errors, etc.) will vary but at least we have something definite to tell us to beware of the results. Another check we used was simply including an additional column of ones in the second matrix of any product, AB; the result will include the row sums of A, which can easily be checked by hand.

A final check on any run of a computer program should always include a test problem at the beginning so that when we examine the output, we can see if the machine performed properly and if the program is doing what it is supposed to do. It can also serve as a final check on ourselves; for example, if we had dimensioned the variables properly according to all the subroutines we may be using. Another check on the output of a computer run comes from a relation that exists between the norms we mentioned in the Introduction, the Frobenius norm, N_F, and the maximum element norm, N_M. It can be shown that $N_F \leq N_M$. Hence, using this relation can give us more information about the validity of the results. We have found all these checks both necessary and effective.

One final note on avoiding input errors. In ill-conditioned problems, if the data are not exact, the results could be very far from correct. Hence, all our data are integers, exactly represented in the machine. This is accomplished by multiplying each matrix by an appropriate scalar, which will be indicated in the next subsection.

The Test Matrices

1. T_{10}^4

This is the fourth power of the 10×10 tridiagonal matrix with -2 on the diagonal, 1 above and below the diagonal, and 0 elsewhere.

$$T = (t_{ij}),\ t_{ij} = \begin{cases} -2, & i = j \\ 1, & |i-j| = 1 \\ 0, & |i-j| \geq 2. \end{cases}$$

$$P(T_{10}^4) \approx (4/\pi^2)^4 \cdot (n^2)^4$$

$$\approx .27E + 07.$$

The elements are all integers.

$$T^{-1} = (b_{ij}),\ b_{ij} = \begin{cases} \dfrac{-i(n-j+1)}{n+1}, & i \leq j \\ b_{ji}, & i > j. \end{cases}$$

2. T_{20}^3

T is the same as above and we just change the dimension to 20 and the power to 3.

$$P(T_{20}^3) \approx (4/\pi^2)^3 \cdot (n^2)^3$$

$$\approx .43E + 07.$$

The elements are all integers.

3. T_{20}^4

T is the same as above and we just change the power to 4.

$$P(T_{20}^4) \approx (4/\pi^2)^4 \cdot (n^2)^4$$

$$\approx .69E + 09.$$

The elements are all integers.

4. A_{100}

This is a 10×10 matrix where $A_k = (1/k)I + J$ and J is the 10×10 matrix of all ones. $P(A_k) = 1 + 10k$. The integer form for use as a test matrix is obviously achieved upon multiplication by k.

$$(kA_k)^{-1} = I - \frac{k}{1+10k} J.$$

$$P(A_{100}) = 1001$$
$$\approx .10E + 04.$$

5. A_{1000}

The same as above except that we change the value of k.

$$P(A_{1000}) = 10001$$
$$\approx .10E + 05.$$

6. A_{10000}

The same as above except that we change the value of k.

$$P(A_{10000}) = 100001$$
$$\approx .10E + 06.$$

7. H_8

H_n is the $n \times n$ Hilbert matrix.

$$H_n = (h_{ij}),\ h_{ij} = (i+j-1)^{-1}$$

and

$$\log_e [P(H_n)] \approx 3.5n.$$

The integer form for use as a test matrix is achieved upon multiplication by the least common multiple of $(1, 2, 3, \ldots, 2n-1)$.

$$H_n^{-1} = (b_{ij})$$

where

$$b_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!}.$$

$$P(H_6) \approx .13E + 10$$

and the multiplier for integer input is 27720.

8. H_8

This is the 8×8 Hilbert matrix.

$$P(H_8) \approx .14E + 13$$

and the scalar multiplier for integer input is 360360.

9. H_{10}

This is the 10×10 Hilbert matrix.

$$P(H_{10}) \approx .16E+16$$

and the scalar multiplier for integer input is 232792560.

The Computer Programs

1. LEQ

A FORTRAN subroutine used to solve the matrix equation $AX=B$ and to evaluate the determinant of A. It was written by Max Goldstein of the AEC Computing and Applied Mathematics Center at the Courant Institute of Mathematical Sciences, New York University. The Gauss elimination method is used. The matrices are normalized row-wise by dividing by the largest element of $A(I, J)$ in that row, then the A matrix is reduced to triangular form by $(N-1)$ transformations using a pivotal condensation process after which $X(I, K)$ is computed by a back-substitution process. This transforms B into X and leaves the product of the diagonal elements as the determinant of A.

2. MATH PACK, GJR

A FORTRAN subroutine, one of the Univac 1108 Math Pack programs, available on the library tapes at the University of Maryland computing center. It solves simultaneous equations, computes a determinant, or inverts a matrix or any combination of the three above by using a Gauss-Jordan elimination technique with column pivoting.

3. MATH PACK, MXHOI

A FORTRAN subroutine, one of the Univac 1108 Math Pack programs, available on the library tapes at the University of Maryland computing center. It improves the accuracy of the inverse of a matrix by an iterative method which has cubic convergence.

4. MIDAS

A FORTRAN and ALGOL package to solve general nonsingular systems of linear algebraic equations, invert matrices, and compute determinants. Error bounds on the solution or inverse are available as an option. It was written by Peter A. Businger of Bell Telephone Laboratories, Inc., Murray Hill, New Jersey. The error bound is a bound on the distance between any element of the true inverse and the corresponding element of the computed inverse (unless the bound equals -1, in which case no bound is available). Gaussian elimination with partial pivoting is used to decompose the $N \times N$ input matrix into the product of a lower and an upper triangular matrix (LU decomposition). The magnitude of intermediate results is estimated; in case of alarming growth the program switches to complete pivoting. When solving a system of equations $Ax=b$, the accuracy of the solution obtained from the triangular systems is improved by iteration; in the case of matrix inversion the iteration is omitted for the sake of computational efficiency. (We note that the error bound is essentially $N(X)N(Y)/[1-N(Y)]$, $Y=I-AX$.)

5. MINV

A FORTRAN subroutine, one of the IBM System/360 Scientific Subroutine Package. It inverts a general matrix by the standard Gauss-Jordan method. The determinant is also calculated. A determinant of zero indicates a singular matrix.

6. OMNITAB, INVERT

OMNITAB is a general-purpose computer program for statistical and numerical analysis developed at the National Bureau of Standards by Joseph Hilsenrath et al. Now available in an ASA FORTRAN version, OMNITAB allows the user to communicate with a computer in an efficient manner by means of simple English sentences. INVERT is the calling word for the matrix inversion routine used and an error bound is given: $N(X)N(Y)/[1-N(Y)]$.

408-680 O - 71 - 5

7. SOLVE

A FORTRAN program by Cleve Moler given in the book *Computer Solution of Linear Algebraic Systems* by George Forsythe and Cleve Moler. It uses Gauss elimination with partial pivoting and has a subroutine IMPRUV, which can be called to improve the solution of a linear algebraic system. Appropriate messages for various kinds of singularity are available. It is presently undergoing some changes to increase efficiency in most FORTRAN systems although these changes should not materially alter the numerical behavior.

8. LSD

This is a program, with automatic error bounding, for solving a system of linear equations. This is equivalent to finding the solution, with error bounds, of the matrix equation $Ax=b$ where A is an $N\times N$ matrix and b and x are $N\times 1$ vectors. The error bounding is done with interval arithmetic (I.A.). The program permits interval input and hence errors may be included which arise from uncertainty in the data. A second entry to LSD is provided for solving several sets of equations with the same A. Thus the program is efficient in finding the inverse of A, with error bounds, by using successively for b the columns of the identity matrix. Output from the program includes an upper and lower bound on the solution vector. Therefore, the number of significant figures in the answer may be simply read from the output. LSD uses two FORTRAN subroutines, a FAP function, and a FAP subroutine in addition to interval arithmetic subroutines coded in FAP. The program requires only standard FORTRAN II software. The double precision arithmetic hardware feature of the IBM 7094 is used, thus restricting the use of LSD, in its present form to this machine. It was written by Eldon Hansen and Roberta Smith.

We intended to include four other programs in our evaluation:

1. STAT PACK, JIM

A FORTRAN subroutine, one of the Univac 1108 STAT-PACK programs, available on the library tape at the University of Maryland computing center. Apparently, this subroutine never functioned properly and the subroutine actually calls GJR of MATH PACK.

2. SPVMTX

A single precision FORTRAN IV program for inverting a matrix or solving a set of linear equations. To a program from the SHARE library (7090–F1 3180INV1 Single Precision Matrix Inversion with Selective Pivoting, written by A. R. Sadaka), Sally T. Peavy, National Bureau of Standards, incorporated accuracy checks. We belatedly discovered that this is the routine used by INVERT of OMNITAB. This was confirmed by the identical outputs of both programs.

3. LEQU

A FORTRAN subroutine to approximate the solution X of the system $AX=B$ of linear algebraic equations, where A is an $N\times N$ matrix and B and X are $N\times M$ matrices. This program is faster but less accurate than subroutine LEQUN, which should normally be preferred. It was written by William Kahan, who was then at the Institute of Computer Science, University of Toronto (presently at the University of California, Berkeley). We found that this program depends heavily on other routines in the University of Toronto 7094 library and we were unable to delete this dependence or acquire the other routines at this time.

4. ORTHO1

An ALGOL subroutine (with iterative improvement) to invert a matrix. It was written by F. L. Bauer and essentially uses Gauss elimination but with a suitably weighted combination of rows

used for elimination instead of a single row. We were unable, at this time, to run the program in ALGOL or to translate it into FORTRAN.

We would like to mention one other program that exists, one which solves equations exactly. It is called SOLVER and was written in FORTRAN by Dr. Morris Newman of the National Bureau of Standards. It employs a congruential method and calculates the exact solution of $Ax=b$ where the elements of A and b must be integers. It will solve systems in which A is a square matrix at most 100×100 and the elements of A and b are numerically less than 10^{20}. This method is not at all sensitive to the condition of A, but it can be time-consuming for large systems. It has successfully inverted the Hilbert matrices of all orders up to and including $n=13$ (a limit imposed by considerations of time). Since our test matrices all have known inverses, it was not necessary to run this program.

The Results

In the following tables we list some of the information that we had available from the execution of each program for the test problems. We rank the programs according to the smallness of the error bound. In some cases, the values of this bound were quite close so the difference in the rank may sometimes be slightly artificial. We did not include MXHOI in the rankings and this will be explained later. For our estimates we used the Frobenius norm and the maximum element norm and we group our results accordingly. In the cases where $N(I-AX)$ was greater than 1, our error bounds do not apply and so we leave that row blank. With regard to the check vector carried throughout the inversion process, we subtracted this from the vector $b=(1, 1, \ldots, 1)^T$ and list the maximum absolute value of this difference. We note that for a check matrix (T_6 with a P-condition number of approximately 15) all the programs performed extremely well, the only difference, if any, being in the last digit. We will discuss the interval arithmetic program of Hansen and Smith separately, immediately after presenting the results of the other programs.

All the calculations, except those for the interval arithmetic program, were performed on the Univac 1108 at the University of Maryland between March 10 and March 22, 1970. In use at this time was version 23.65.74:N of the EXEC 8 system.

TABLE 1. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.21 E−02	.13 E−01	.29 E+02	.30 E+03	.12 E−02	.23 E+01	2
MATH PAK	GJR	.16 E−01	.69 E+00	.12 E+04	.51 E+05	.86 E−01	.34 E−01	6
	MXHOI	.21 E−02	.25 E−01	.66 E+02	.60 E+03	.29 E−02	.35 E+04	
MIDAS		.83 E−02	.29 E−01	.15 E+03	.68 E+03	.65 E−02	.12 E+00	4
MINV		.87 E−02	.29 E−01	.15 E+03	.69 E+03	.66 E−02	.27 E−01	4
OMNITAB (INVERT)		.12 E−01	.12 E+01				.46 E−01	7
SOLVE	no IMPRUV	.35 E−02	.86 E−02	.76 E+02	.20 E+03	.33 E−02	.73 E+00	3
	IMPRUV	.75 E−08	.81 E−02	.15 E−03	.19 E+03	.65 E−08	.94 E−02	1

Matrix T_{10}^4
Condition number .27 E+07
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y=I-AX$

$$N(A^{-1}-X)\leq\frac{N(XY)}{1-N(Y)}\leq\frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})}\leq\frac{N(XY)}{1-N(Y)}\cdot\frac{1+N(Y)}{N(X)}$$

TABLE 2. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.20 E−02	.13 E−01	.22 E+03	.12 E+04	.25 E−02	.93 E+01	3
MATH PAK	GJR	.12 E−02	.66 E+01				.20 E−01	6
	MXHOI	.41 E−02	.31 E−01	.31 E+03	.28 E+04	.35 E−02	.65 E+04	
MIDAS		.38 E−02	.20 E−01	.24 E+03	.18 E+04	.27 E−02	.57 E+01	4
MINV		.45 E−02	.31 E−01	.36 E+03	.29 E+04	.41 E−02	.23 E−01	5
OMNITAB (INVERT)		.17 E−02	.70 E+01				.36 E−01	6
SOLVE	no IMPRUV	.17 E−02	.15 E−01	.18 E+03	.14 E+04	.21 E−02	.58 E+01	2
	IMPRUV	.75 E−08	.99 E−02	.60 E−03	.90 E+03	.68 E−08	.10 E−01	1

Matrix T^3_{20}
Condition number .43 E+07
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 3. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.47 E+01	.78 E+01				.22 E+05	
MATH PAK	GJR	.71 E+00	.36 E+04				.31 E+01	
	MXHOI		overflow					
MIDAS		.51 E+00	.28 E+01				.15 E+04	
MINV		.69 E+00	.32 E+01				.21 E+01	
OMNITAB (INVERT)		.37 E+00	.98 E+04				.45 E+01	
SOLVE	no IMPRUV	.71 E+00	.10 E+01				.30 E+04	
	IMPRUV	.13 E−02	.18 E+01				.14 E+02	

Matrix T^4_{20}
Condition number .69 E+09
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 4. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.75 $E-06$	.17 $E-04$	.12 $E-05$	.50 $E-04$	.39 $E-06$	.20 $E-04$	7
MATH PAK	GJR	.26 $E-05$	.42 $E-04$	.84 $E-06$	.13 $E-03$	.28 $E-06$	.41 $E-04$	6
	MXHOI	.26 $E-05$	.42 $E-04$	.84 $E-06$	.13 $E-03$	.28 $E-06$	.41 $E-04$	
MIDAS		.86 $E-05$	.66 $E-04$	.13 $E-06$	.20 $E-03$	.43 $E-07$	.20 $E-04$	3
MINV		.30 $E-07$	.36 $E-04$	.39 $E-07$	.11 $E-03$	.13 $E-07$	.19 $E-04$	2
OMNITAB (INVERT)		.25 $E-05$	.41 $E-04$	.83 $E-06$	.12 $E-03$	.28 $E-06$	.42 $E-04$	5
SOLVE	no IMPRUV	.26 $E-05$	.15 $E-04$	.82 $E-06$	.44 $E-04$	.27 $E-06$	.18 $E-04$	4
	IMPRUV	.75 $E-08$	.72 $E-05$	.14 $E-07$	.22 $E-04$	.47 $E-08$	.72 $E-05$	1

Matrix A_{108}
Condition number .10 $E+04$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 5. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.34 $E-04$	.18 $E-03$	.13 $E-04$	.54 $E-03$	.44 $E-05$	.18 $E-03$	7
MATH PAK	GJR	.12 $E-03$	.33 $E-03$	.65 $E-05$	.10 $E-02$	.22 $E-05$	.34 $E-03$	4
	MXHOI	.12 $E-03$	.33 $E-03$	.65 $E-05$	.10 $E-02$	.22 $E-05$	.34 $E-03$	
MIDAS		.67 $E-04$	.30 $E-03$	.89 $E-07$	.91 $E-03$	.30 $E-07$	.24 $E-03$	3
MINV		.30 $E-07$	.29 $E-03$	.36 $E-07$	.86 $E-03$	.12 $E-07$	.18 $E-03$	2
OMNITAB (INVERT)		.14 $E-04$	.43 $E-03$	.65 $E-05$	.13 $E-02$	.22 $E-05$	.43 $E-03$	4
SOLVE	no IMPRUV	.14 $E-04$	.15 $E-03$	.65 $E-05$	.46 $E-03$	.22 $E-05$	.16 $E-03$	4
	IMPRUV	.15 $E-07$	.84 $E-04$	.94 $E-08$	.25 $E-03$	.31 $E-08$	.84 $E-04$	1

Matrix A_{1000}
Condition number .10 $E+05$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 6. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.42 $E-03$	.13 $E-02$	.30 $E-03$	.39 $E-02$	.10 $E-03$	.18 $E-02$	7
MATH PAK	GJR	.22 $E-06$	.29 $E-02$	.90 $E-04$	.88 $E-02$	.30 $E-04$	.30 $E-02$	4
	MXHOI	.80 $E-06$	.54 $E-02$	.15 $E-03$	.16 $E-01$	.51 $E-04$	.31 $E+01$	
MIDAS		.11 $E-02$	.27 $E-02$	.29 $E-06$	.81 $E-02$	.96 $E-07$	.17 $E-02$	3
MINV		.30 $E-07$	.32 $E-02$	.36 $E-07$	.95 $E-02$	.12 $E-07$	.16 $E-02$	2
OMNITAB (INVERT)		.60 $E-07$	.41 $E-02$	.90 $E-04$	.12 $E-01$	.30 $E-04$	.41 $E-02$	4
SOLVE	no IMPRUV	.30 $E-07$	.16 $E-02$	.90 $E-04$	.48 $E-02$	.30 $E-04$	.20 $E-02$	4
	IMPRUV	.15 $E-07$	.48 $E-03$	.10 $E-07$	.14 $E-02$	.35 $E-08$	.48 $E-03$	1

Matrix A_{10000}
Condition number .10 $E+06$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 7. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.20 $E-01$	.19 $E-01$	.92 $E+00$	.63 $E+01$	.28 $E-02$	.47 $E-01$	2
MATH PAK	GJR	.30 $E-02$	.30 $E-01$	.22 $E+01$	.10 $E+02$	.67 $E-02$	.31 $E-01$	7
	MXHOI	.11 $E-01$	.34 $E-01$	.12 $E+01$	.12 $E+02$	.36 $E-02$	.56 $E+03$	
MIDAS		.10 $E-01$	.28 $E+00$	.18 $E+01$	.13 $E+03$	.67 $E-02$	.41 $E-01$	6
MINV		.86 $E-02$	.56 $E-01$	.11 $E+01$	.20 $E+02$	.36 $E-02$	.56 $E-01$	3
OMNITAB (INVERT)		.10 $E-01$	.40 $E-01$	.17 $E+01$	.14 $E+02$	.52 $E-02$	.36 $E-01$	5
SOLVE	no IMPRUV	.16 $E-01$	.18 $E-01$	.16 $E+01$	.61 $E+01$	.50 $E-02$	.65 $E-01$	4
	IMPRUV	.00	.96 $E-02$	.29 $E-05$	.32 $E+01$	.88 $E-08$	.96 $E-02$	1

Matrix H_8
Condition number .13 $E+10$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 8. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ			overflow					
MATH PAK	GJR	.93 $E+00$	.20 $E+02$				.89 $E+01$	
	MXHOI		overflow					
MIDAS		.77 $E+01$	.16 $E+03$				.10 $E+03$	
MINV		.22 $E+02$	.26 $E+02$				.23 $E+02$	
OMNITAB (INVERT)		.67 $E+00$	.66 $E+01$				.52 $E+01$	
SOLVE	no IMPRUV	.19 $E+01$	.20 $E+01$				.26 $E+02$	
	IMPRUV	.16 $E+02$	.10 $E+03$				.35 $E+03$	

Matrix H_8
Condition number .14 $E+13$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 9. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ			overflow					
MATH PAK	GJR	.79 $E+01$	.14 $E+04$				.67 $E+02$	
	MXHOI		overflow					
MIDAS		.89 $E+01$	.35 $E+02$				.96 $E+01$	
MINV		.18 $E+02$	.24 $E+02$				.22 $E+04$	
OMNITAB (INVERT)		.17 $E+02$	.76 $E+02$				.15 $E+02$	
SOLVE	no IMPRUV	.27 $E+02$	.39 $E+02$				.16 $E+04$	
	IMPRUV	.47 $E+03$	.59 $E+03$				.85 $E+05$	

Matrix H_{10}
Condition number .16 $E+16$
Norm Frobenius

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 10. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.21 E−02	.44 E−01	.53 E+02	.19 E+04	.13 E−02	.66 E+01	5
MATH PAK	GJR	.16 E−01	.18 E+01				.11 E+00	6
	MXHOI	.21 E−02	.91 E−01	.22 E+03	.42 E+04	.58 E−02	.86 E+04	
MIDAS		.83 E−02	.13 E+00	.29 E+03	.59 E+04	.79 E−02	.47 E+00	4
MINV		.87 E−02	.70 E−01	.27 E+03	.31 E+04	.71 E−02	.92 E−01	3
OMNITAB (INVERT)		.12 E−01	.26 E+01				.12 E+00	6
SOLVE	no IMPRUV	.35 E−02	.35 E−01	.14 E+03	.15 E+04	.35 E−02	.23 E+01	2
	IMPRUV	.75 E−08	.27 E−01	.46 E−03	.11 E+04	.11 E−07	.29 E−01	1

Matrix T_{10}^{4}

Condition number .27 E+07

Norm Maximum element

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 11. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.20 E−02	.68 E−01	.45 E+03	.12 E+05	.28 E−02	.30 E+02	3
MATH PAK	GJR	.12 E−02	.23 E+02				.76 E−01	6
	MXHOI	.41 E−02	.13 E+00	.96 E+03	.25 E+05	.63 E−02	.19 E+05	
MIDAS		.38 E−02	.77 E−01	.48 E+03	.14 E+05	.30 E−02	.28 E+02	4
MINV		.45 E−02	.12 E+00	.75 E+03	.24 E+05	.50 E−02	.11 E+00	5
OMNITAB (INVERT)		.17 E−02	.17 E+02				.12 E+00	6
SOLVE	no IMPRUV	.17 E−02	.68 E−01	.37 E+03	.12 E+05	.23 E−02	.19 E+02	2
	IMPRUV	.75 E−08	.42 E−01	.22 E−02	.74 E+04	.13 E−07	.42 E−01	1

Matrix T_{10}^{3}

Condition number .43 E+07

Norm Maximum element

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 12. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEG		.47 E+01	.31 E+02				.81 E+05	
MATH PAK	GJR	.71 E+00	.11 E+05				.11 E+02	
	MXHOI		overflow					
MIDAS		.51 E+00	.12 E+02				.73 E+04	
MINV		.69 E+00	.83 E+01				.79 E+01	
OMNITAB (INVERT)		.37 E+00	.30 E+05				.16 E+02	
SOLVE	no IMPRUV	.71 E+00	.38 E+01				.15 E+05	
	IMPRUV	.13 E−02	.73 E+01				.44 E+02	

Matrix T^3_{20}

Condition number .69 E+09

Norm Maximum element

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

*Difference from 1.

TABLE 13. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.73 E−06	.29 E−04	.10 E−04	.26 E−03	.11 E−05	.38 E−04	7
MATH PAK	GJR	.26 E−05	.69 E−04	.76 E−05	.62 E−03	.84 E−06	.66 E−04	6
	MXHOI	.26 E−05	.69 E−04	.76 E−05	.62 E−03	.84 E−06	.66 E−04	
MIDAS		.86 E−05	.17 E−03	.41 E−06	.16 E−02	.46 E−07	.33 E−04	3
MINV		.30 E−07	.11 E−03	.12 E−06	.98 E−03	.13 E−07	.20 E−04	2
OMNITAB (INVERT)		.25 E−05	.57 E−04	.74 E−05	.51 E−03	.82 E−06	.58 E−04	5
SOLVE	no IMPRUV	.26 E−05	.31 E−04	.73 E−05	.28 E−03	.82 E−06	.25 E−04	4
	IMPRUV	.75 E−08	.76 E−05	.42 E−07	.69 E−04	.47 E−08	.76 E−05	1

Matrix A_{100}

Condition number .10 E+04

Norm Maximum element

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

*Difference from 1.

TABLE 14. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.34 E−04	.34 E−03	.58 E−04	.31 E−02	.64 E−05	.35 E−03	7
MATH PAK	GJR	.12 E−03	.57 E−03	.59 E−04	.52 E−02	.65 E−05	.68 E−03	4
	MXHOI	.12 E−03	.57 E−03	.59 E−04	.52 E−02	.65 E−05	.68 E−03	
MIDAS		.67 E−04	.80 E−03	.54 E−06	.72 E−02	.60 E−07	.37 E−03	3
MINV		.30 E−07	.85 E−03	.93 E−07	.77 E−02	.10 E−07	.18 E−03	2
OMNITAB (INVERT)		.14 E−04	.63 E−03	.58 E−04	.57 E−02	.65 E−05	.59 E−03	4
SOLVE	no IMPRUV	.14 E−04	.29 E−03	.58 E−04	.26 E−02	.65 E−05	.27 E−03	4
	IMPRUV	.15 E−07	.88 E−04	.18 E−07	.79 E−03	.20 E−08	.88 E−04	1

Matrix A_{1000}
Condition number .10 E+05
Norm Maximum element

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 15. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.42 E−03	.34 E−02	.21 E−02	.31 E−01	.24 E−03	.30 E−02	7
MATH PAK	GJR	.22 E−06	.59 E−02	.82 E−03	.53 E−01	.91 E−04	.65 E−02	5
	MXHOI	.80 E−06	.96 E−02	.70 E−03	.87 E−01	.79 E−04	.63 E+01	
MIDAS		.11 E−02	.60 E−02	.12 E−05	.54 E−01	.13 E−06	.25 E−02	3
MINV		.30 E−07	.98 E−02	.11 E−06	.89 E−01	.12 E−07	.16 E−02	2
OMNITAB (INVERT)		.60 E−07	.61 E−02	.82 E−03	.55 E−01	.91 E−04	.62 E−02	5
SOLVE	no IMPRUV	.30 E−07	.29 E−02	.81 E−03	.26 E−01	.91 E−04	.40 E−02	4
	IMPRUV	.15 E−07	.49 E−03	.31 E−07	.45 E−02	.35 E−08	.49 E−03	1

Matrix A_{10000}
Condition number .10 E+06
Norm Maximum element

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

TABLE 16. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ		.20 E−01	.57 E−01	.26 E+01	.58 E+02	.29 E−02	.12 E+00	2
MATH PAK	GJR	.30 E−02	.12 E+00	.68 E+01	.14 E+03	.80 E−02	.97 E−01	6
	MXHOI	.11 E−01	.11 E+00	.32 E+01	.12 E+03	.37 E−02	.12 E+04	
MIDAS		.10 E−01	.86 E+00	.27 E+02	.61 E+04	.52 E−01	.16 E+00	7
MINV		.86 E−02	.19 E+00	.37 E+01	.23 E+03	.46 E−02	.22 E+00	3
OMNITAB (INVERT)		.10 E−01	.91 E−01	.51 E+01	.96 E+02	.58 E−02	.12 E+00	5
SOLVE	no IMPRUV	.16 E−01	.61 E−01	.49 E+01	.62 E+02	.54 E−02	.13 E+00	4
	IMPRUV	.00	.23 E−01	.11 E−04	.22 E+02	.11 E−07	.23 E−01	1

Matrix H_8
Condition number .13 E+10
Norm Maximum element

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

*Difference from 1.

TABLE 17. *Summary of results*

Program		Max ELT* of check vector	$N(I-AX)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ			overflow					
MATH PAK	GJR	.93 E+00	.67 E+02				.25 E+02	
	MXHOI		overflow					
MIDAS		.77 E+01	.49 E+03				.67 E+03	
MINV		.22 E+02	.81 E+02				.80 E+02	
OMNITAB (INVERT)		.67 E+00	.29 E+02				.16 E+02	
SOLVE	no IMPRUV	.19 E+01	.80 E+01				.66 E+02	
	IMPRUV	.16 E+02	.26 E+03				.12 E+04	

Matrix H_8
Condition number .14 E+13
Norm Maximum element

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

*Difference from 1.

TABLE 18. *Summary of results*

Program		Max ELT* of check vector	$N(I-AY)$	$\frac{N(XY)}{1-N(Y)}$	$\frac{N(X)N(Y)}{1-N(Y)}$	Relative error	$N(I-XA)$	Rank
LEQ			overflow					
MATH PAK	GJR	.79 $E+01$	.58 $E+04$				.26 $E+03$	
	MXHOI		overflow					
MIDAS		.89 $E+01$	.12 $E+03$				.61 $E+02$	
MINV		.18 $E+02$	.12 $E+03$				.94 $E+02$	
OMNITAB (INVERT)		.17 $E+02$	.20 $E+03$				.79 $E+02$	
SOLVE	no IMPRUV	.27 $E+02$	.13 $E+03$				.40 $E+04$	
	IMPRUV	.47 $E+03$	.24 $E+04$				.25 $E+06$	

Matrix H_{10}
Condition number .16 $E+16$
Norm Maximum element

*Difference from 1.

X = Approximate inverse

$Y = I - AX$

$$N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$$

$$\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$$

For the program LSD, using interval arithmetic, we found excellent results for all the matrices except H_8, H_{10}, and T^4_{20}. By excellent results we mean the upper and lower values given for each element of the inverse agreed to 7 or 8 digits with each other and with the exact value. For H_8, the spread between the upper and lower values was quite different. A typical situation is given in the first row below, and the worst spread in the second row.

A^{-1} Lower	Exact	A^{-1} Upper
.12533643 $E-03$	.17760000 $E-03$	.22986282 $E-03$
$-$.16101655 $E+01$	$-$.80000000 $E+00$	$+$.10338515 $E-01$

In view of the fact that none of the other programs solved this problem, these results are not bad.

For the 10×10 Hilbert matrix, the following error message was printed out: SINGULAR MATRIX INCLUDED IN INPUT MATRIX. (We note that SOLVER produced the exact inverse in both of these cases.)

For T^4_{20} we got results but the spread was so large that they are meaningless. One of the better values was from .26055424 $E+05$ to .56997728 $E+05$ but a more common spread would be similar to the following:

$-$.14807597 $E+07$ to $+$.18754179 $E+07$

or $-$.60322834 $E+08$ to $+$.60828502 $E+08$.

These certainly are not very satisfactory. Let us now turn our attention to the other programs.

Evaluation and Discussion

It is obvious that SOLVE with iterative improvement performed far better than any of the others. The excellent results that are achieved by using this added feature certainly seem to warrant the extra work involved. In each case where it did not succeed in inverting the matrix (H_8, H_{10}, T^4_{20}), the following error message was printed out: ITERATION DID NOT CONVERGE. MATRIX IS NEARLY

SINGULAR. This is most important because it tells us our answers may not be correct. LEQ had floating point overflow, MIDAS had the error bound of -1, and OMNITAB had a negative bound, each of which also indicates doubtful results. MINV and GJR gave no indication that the results might not be correct—a very serious defect. GJR does, in fact, have an error return but it is only for overflow and this is not sufficient. The zero determinant criterion of MINV is also not adequate. Knowing that the condition number is as high as it is for these two matrices, we expected trouble. But for T^4_{20} whose condition number is fairly large but whose largest element is 70, we did not anticipate serious difficulty and the message of singularity was needed. In this case, however, LEQ did not have any floating point overflow message.

If we did not know ahead of time that the matrix was poorly conditioned, we could use the iterative improvement scheme if for no other reason than to inform us of doubtful results. In solving a single system of equations where we may not calculate an approximate inverse (from which a residual matrix whose norm being greater than 1 can indicate trouble in spite of the program output), the improvement scheme seems an excellent way of indicating possible poor results.

In the cases where the inversions were successfully accomplished (that is, at least $N(I-AX)$ was less than 1), it seems that MINV, SOLVE (no IMPRUV), and MIDAS were more successful than the others, excluding IMPRUV of course. It is interesting to note that the only program (besides IMPRUV) which had the norm of *both* residuals less than one was MINV. This is certainly a desirable situation. It also points out the large discrepancy that can exist between the two residual matrices and the fact that the righthand residual, $I-AX$, is not always smaller. Hence, a bound on the error matrix can sometimes be available by using the other residual. None of the programs seem to take this possibility into consideration. If we have gone to the trouble of calculating $I-AX$ and get a negative bound, it might be worthwhile to calculate $I-XA$ and see if a bound can be found. This will also indicate the kind of inverse we have. It is still, of course, up to the originator of the problem to decide if this is necessary.

The vagaries of numerical calculations are also clearly indicated by these results. We have cases in the same problem where $N(I-AX) > N(I-XA)$ for one program and the opposite is true for another program; for the same program, we may have $N(I-AX) > N(I-XA)$ for one problem and the opposite for another problem. Hence, one must always be extremely careful in doing numerical work.

The poor and erratic performance of MXHOI led to a brief investigation of this improvement procedure. It never really improved the results and in some cases (A_{10000}) made them worse. We found that no double precision was used at all and, as we indicated in section 3, for a scheme of cubic convergence, multiprecision may be needed, let along double precision. The performance of MXHOI certainly bears this out! The results are completely unreliable and this is why we did not even consider this routine as far as ranking the various programs was concerned.

Finally, we would like to discuss briefly the merits of the bound $N(XY)/[1-N(Y)]$. We have observed in section 1 that using this bound would probably give better results for the relative error than the simpler expression $N(Y)$, where Y is either residual matrix. We note that this is in fact true in every one of our problems where the inversion was successful.

Another and much more important fact is the following. We have attempted in our first part to derive a practical bound for $N(A^{-1}-X)$; i.e., one that is realistic and not misleading. It is clear that using $N(XY)$ is better than $N(X)N(Y)$ —there is almost always at least a factor of 10 involved in the improvement. This is certainly worthwhile. But just how close to the actual error is our bound? This question can be partially answered by looking at the lower bound, $N(E)/2N(A)$. This was not included in the tables because the only time it was close to the upper bound was for SOLVE with IMPRUV, which certainly indicates both the effectiveness of this scheme and the reality of the upper bound. A more complete answer comes from calculating the actual error. We chose H_6 and T^3_{20} as being of the more difficult type and list the results in the following tables. The comparisons are extremely satisfying.

TABLE 19. *Norms of actual difference*

Program		Maximum Element Norm			Frobenius Norm		
		$\frac{N(XY)}{1-N(Y)}$	$N(A^{-1}-X)$ (actual)	$\frac{N(X)N(Y)}{1-N(Y)}$	$\frac{N(XY)}{1-N(Y)}$	$N(A^{-1}-X)$ (actual)	$\frac{N(X)N(Y)}{1-N(Y)}$
LEQ		.26 E+01	.25 E+01	.58 E+02	.92 E+00	.91 E+00	.63 E+01
MATH PAK	GJR	.68 E+01	.59 E+01	.14 E+03	.22 E+01	.21 E+01	.10 E+02
	MXHOI						
MIDAS		.27 E+02	.37 E+01	.61 E+04	.18 E+01	.13 E+01	.13 E+03
MINV		.37 E+01	.30 E+01	.23 E+03	.11 E+01	.11 E+01	.36 E+02
OMNITAB (INVERT)		.51 E+01	.46 E+01	.96 E+02	.17 E+01	.16 E+01	.14 E+02
SOLVF	no IMPRUV	.49 E+01	.46 E+01	.62 E+02	.16 E+01	.16 E+01	.61 E+01
	IMPRUV	.11 E−04	.60 E−06	.22 E+02	.29 E−05	.14 E−06	.32 E+01

Matrix H_6 X = Approximate inverse $N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$

Condition number .13 E+10 $Y = I - AX$ $\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$

TABLE 20. *Norms of actual difference*

Program		Maximum Element Norm			Frobenius Norm		
		$\frac{N(XY)}{1-N(Y)}$	$N(A^{-1}-X)$ (actual)	$\frac{N(X)N(Y)}{1-N(Y)}$	$\frac{N(XY)}{1-N(Y)}$	$N(A^{-1}-X)$ (actual)	$\frac{N(X)N(Y)}{1-N(Y)}$
LEQ		.45 E+03	.42 E+03	.12 E+05	.22 E+03	.22 E+03	.12 E+04
MATH PAK	GJR**	.40 E+03	.37 E+03	.14 E+05	.19 E+03	.19 E+03	.18 E+04
	MXHOI						
MIDAS		.48 E+03	.44 E+03	.14 E+05	.24 E+03	.23 E+03	.18 E+04
MINV		.75 E+03	.66 E+03	.24 E+05	.36 E+03	.35 E+03	.29 E+04
OMNITAB (INVERT)**		.39 E+03	.36 E+03	.24 E+05	.19 E+03	.18 E+03	.34 E+04
SOLVE	nó IMPRUV						
	IMPRUV	.22 E−02	.12 E−02	.74 E+04	.60 E−03	.36 E−03	.90 E+03

Matrix T_{10}^3 X = Approximate inverse $N(A^{-1}-X) \leq \frac{N(XY)}{1-N(Y)} \leq \frac{N(X)N(Y)}{1-N(Y)}$

Condition number .43 E+07 $Y = I - AX$ $\frac{N(A^{-1}-X)}{N(A^{-1})} \leq \frac{N(XY)}{1-N(Y)} \cdot \frac{1+N(Y)}{N(X)}$

**$Y = I - XA$

In conclusion, we would like to make the following observations:

1. Some indication of the possibility of poor results should be included in the output of every program.

2. Some indication of the accuracy of the results should be included in the output of every program.

3. SOLVE with iterative improvement is certainly an excellent program.

4. Interval arithmetic was quite effective but not in every case.

5. The use of $\frac{N(XY)}{1-N(Y)}$ gives a realistic (practical) upper bound on the error of a computed inverse, can be used with either residual matrix and would indicate the possibility of poor results if the norm of each residual matrix were greater than one.

I wish to thank the many people who encouraged me throughout my pursuit of this final step, especially Morris Newman, who made it all possible.

I am grateful to the Lewis and Rosa Strauss Memorial Fund for the financial aid which made the computing facilities of the University of Maryland available to me.

7. References

[1] Fox, L., An Introduction to Numerical Linear Algebra (Oxford Univ. Press, New York, N.Y., 1965), p. 16.
[2] Tompkins, C., and Wilson, W., Jr., Elementary Numerical Analysis (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969), p. 111.
[3] Wampler, R. H., J. Res. Nat. Bur. Stand. (U.S.), **73B** (Math. Sci.), No. 2, 59 (Apr.–June 1969).
[4] Noble, B., Applied Linear Algebra (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969), p. 258.
[5] Belostotskii, A. Y., USSR Comp. Math. Math. Phys. **5**, 151 (1965).
[6] Turing, A. M., Quart J. Mech. Appl. Math. **1**, 301 (1948).
[7] Wilkinson, J. H., Rounding Errors in Algebraic Processes (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963), p. 126.
[8] Albasiny, E. L., Error in digital solution of linear problems, in Error in Digital Computation, Ed. L. B. Rall (John Wiley & Sons, Inc., New York, N.Y., 1965), Vol. 1, p. 171.
[9] Householder, A. S., The Theory of Matrices in Numerical Analysis (Blaisdell Publ. Co., New York, N.Y., 1965), p. 91.
[10] Kahan, W., Can. Math. Bull. **9**, 785 (1966).
[11] Moler, C., J. Assoc. Comp. Mach. **14**, 320 (1967).
[12] Forsythe, G., and Moler, C., Computer Solutions of Linear Algebraic Systems (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1967), p. 110.
[13] Forsythe, G., SIAM Rev. **9**, 504 (1967).
[14] Newman, M., and Todd, J., J. SIAM **6**, 467 (1958).

Selected Bibliography

Books

Faddeev, D. K., and Faddeeva, V. N., Computational Methods of Linear Algebra (San Francisco, Calif., W. H. Freeman & Co., 1963).
Forsythe, George, and Moler, Cleve B., Computer Solution of Linear Algebraic Systems (Englewood Cliffs, N.J., Prentice-Hall, Inc., 1967).
Fox, L., An Introduction to Numerical Linear Algebra (New York, N.Y., Oxford Univ. Press, 1965).
Hildebrand, F. B., Introduction to Numerical Analysis (New York, N.Y., McGraw-Hill Book Co., Inc., 1956).
Householder, Alston S., The Theory of Matrices in Numerical Analysis (New York, N.Y., Blaisdell Publ. Co., 1965).
Kelly, Louis G., Handbook of Numerical Methods and Applications (Reading, Penna., Addison-Wesley Publ. Co., 1967).
Noble, Ben, Applied Linear Algebra (Englewood Cliffs, N.J., Prentice-Hall, Inc., 1969).
Perlis, Sam, Theory of Matrices (Cambridge, Mass., Addison-Wesley Publ. Co., Inc., 1952).
Rall, L. B. (ed.), Error in Digital Computation, vol. I (New York, N.Y., John Wiley & Sons, Inc., 1965).
Ralston, Anthony, and Wilf, Herbert S., Mathematical Methods for Digital Computers, vol. I (New York, N.Y., John Wiley & Sons, Inc., 1962).
Ralston, Anthony, and Wilf, Herbert S., Mathematical Methods for Digital Computers, vol. II (New York, N.Y., John Wiley & Sons, Inc., 1967).
Todd, John (ed.), Survey of Numerical Analysis (New York, N.Y., McGraw-Hill Book Co., Inc., 1962).

Tompkins, Charles B., and Wilson, Walter L., Elementary Numerical Analysis (Englewood Cliffs, N.J., Prentice-Hall, Inc., 1969).

Westlake, Joan R., A Handbook of Numerical Matrix Inversion and Solution of Linear Equations (New York, N.Y., John Wiley & Sons, Inc., 1968).

Wilkinson, J. H., Rounding Errors in Algebraic Processes (Englewood Cliffs, N.J., Prentice-Hall, Inc., 1963).

Wilkinson, J. H., The Algebraic Eigenvalue Problem (Oxford, England, Clarendon Press, 1965).

Articles

Albasiny, E. L., Error in digital solution of linear problems, Error in Digital Computation, ed. L. B. Rall (John Wiley & Sons, Inc., New York, 1965), **1**, 131–184.

Ashenhurst, R. L., Techniques for automatic error monitoring and control, Error in Digital Computation, ed. L. B. Rall (John Wiley & Sons, Inc., New York, 1965), **1**, 43–60.

Ashenhurst, R. L., and Metropolis, N., Unnormalized floating point, J. ACM **6**, No. 3, 415–428 (July 1959).

Bauer, F. L., Elimination with weighted row combinations for solving linear equations and least squares problems, Numer. Math. **7**, No. 4, 338–352 (1965).

Belostotskii, A. Y., Estimation of the quality of approximate solutions of a system of linear algebraic equations, USSR Comp. Math. Math. Phys. **5**, No. 1, 151–154 (1965).

Chartres, B., Automatically controlled precision calculations, J. ACM **13**, 386–403 (July 1966).

Chartres, B., and Geuder, J., Computable error bounds for direct solution of linear equations, J. ACM **14**, No. 1, 63–71 (Jan. 1967).

Descloux, J., Note on round-off errors in iterative processes, Math. Comp. **17**, No. 81, 18–27 (Jan. 1963).

Forsythe, G., Today's computational methods of linear algebra, SIAM Review **9**, No. 3, 489–515 (July 1967).

Golstein, M., Significance arithmetic on a digital computer, Comm. ACM **6**, 111–117 (1963).

Gray, H., and Harrison, C., Jr., Normalized floating point with index of significance, Proc. Eastern Joint Computer Conf. (1959), 244–248.

Kahan, W., Numerical linear algebra, Can. Math. Bull. **9**, No. 6, 757–798 (1966).

Lietzke, M. H., Stroughton, R. W., Lietzke, M. P., A comparison of several methods for inverting large symmetric positive definite matrices, Math. Comp. **18**, No. 87, 449–456 (July 1964).

Loizou, G., An empirical estimate of the relative error of the computed solution $\bar{x}$ of $Ax = b$, Comp. J. **11**, No. 1, 91–94 (May 1968).

Longley, J., An appraisal of least squares programs for the electronic computer from the point of view of the user, J. AM. Stat. Assoc. **62**, No. 319, 819–841 (Sept. 1967).

Moler, C., Iterative refinement in floating point, J. ACM **14**, No. 2, 316–321 (Apr. 1967).

Moore, R., The automatic analysis and control of error in digital computing based on the use of interval numbers, Error in Digital Computation, ed. L. B. Rall (John Wiley & Sons, Inc., New York, 1965), **1**, 61–130.

von Neumann, J., and Goldstine, H., Numerical inverting of matrices of high order, Bull. Am. Math. Soc. **53**, 1021–1099 (1947).

Newman, Morris, Matrix computations, Survey of Numerical Analysis, ed. John Todd (McGraw-Hill Book Co., Inc., New York, 1962), 222–254.

Newman, Morris, Solving equations exactly, J. Res. Nat. Bur. Stand. (U.S.), **71B** (Math. & Math. Phys.), No. 4, 171–179 (Oct.–Dec. 1967).

Newman, M., and Todd, J., The evaluation of matrix inversion programs, J. SIAM **6**, No. 4, 466–476 (Dec. 1958).

Oettli, W., and Prager, W., Compatibility of approximate solutions of linear equations with given error bounds for the coefficients and right hand sides, Numer. Math. **6**, No. 5, 405–409 (1964).

Shroeder, J., Computing error bounds in solving linear systems, Math. Comp. **16**, No. 79, 323–337 (July 1962).

Todd, J., The problem of error in digital computation, Error in Digital Computation, ed. L. B. Rall (John Wiley & Sons, Inc., New York, 1965), **1**, 3–42.

Turing, A. M., Rounding-off errors in matrix processes, Quart. J. Mech. Appl. Math. **1**, 287–308 (1948).

Wampler, R., An evaluation of linear least squares computer programs, J. Res. Nat. Bur. Stand. (U.S.), **73B** (Math. Sci.), No. 2, 59–90 (Apr.–June 1969).

(Paper No. 74B4–336)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

On the Singular Values of a Product of Matrices*

William Watkins**

(September 18, 1970)

The purpose of this note is to give necessary and sufficient conditions for the singular values of a product of matrices to be equal to certain products of their singular values. We then analyze the case of equality in a matrix inequality of Ostrowski.

Key words: Matrices; singular values.

The purpose of this note is to give necessary and sufficient conditions for the singular values of a product of matrices to be equal to certain products of their singular values. We then analyze the case of equality in a matrix inequality of Ostrowski.

The singular values of an n-square complex matrix X are the positive square roots of the eigenvalues of X^*X, where X^* is the conjugate transpose of X. Denote the singular values of X by $\alpha_1(X), \ldots, \alpha_n(X)$, arranged so that $\alpha_1(X) \geq \ldots \geq \alpha_n(X) > 0$ (all matrices are assumed to be nonsingular). Let A and B be n-square complex matrices and let $A = UH$, $B = VK$ be the polar factorizations of A and B. In the factorizations U and V are unitary matrices and H and K are positive-definite hermitian matrices.

THEOREM 1: *Let* k *be a positive integer less than* n. *Then*

$$\alpha_i(\mathrm{AB}) = \alpha_i(\mathrm{A})\alpha_i(\mathrm{B}), \quad \textit{for} \quad 1 \leq \mathrm{i} \leq \mathrm{k} \tag{1}$$

if and only if there exists a unitary matrix W *such that*

$$\mathrm{W^*V^*HVW} = \textit{diag}\ (\alpha_1(\mathrm{A}), \ldots, \alpha_k(\mathrm{A})) \dot{+} \mathrm{T}_1. \tag{2}$$

and

$$\mathrm{W^*KW} = \textit{diag}\ (\alpha_1(\mathrm{B}), \ldots, \alpha_k(\mathrm{B}) \dot{+} \mathrm{T}_2$$

where T_1 *and* T_2 are $(\mathrm{n}-\mathrm{k})$*-square matrices.*

PROOF: Since the singular values of AB are the same as the singular values of $(V^*HV)K$, it suffices to prove this theorem for the case where A and B are positive-definite hermitian matrices.

If there is a unitary matrix W satisfying

$$W^*AW = \mathrm{diag}\ (\alpha_1(A), \ldots, \alpha_k(A)) \dot{+} T_1$$

and

$$W^*BW = \mathrm{diag}\ (\alpha_1(B), \ldots, \alpha_k(B)) \dot{+} T_2 \tag{3}$$

then

$$\alpha_i(AB) = \alpha_i(A)\alpha_i(B) \qquad 1 \leq i \leq k.$$

AMS subject classification. Primary 15A18

*An invited paper.

**Present address: San Fernando Valley State College, Northridge, Calif. 91324

406-680 O - 71 - 6

We use an induction argument on the size of A and B, n, to show that condition (1) implies there is a unitary matrix W satisfying (3). Let $|v| = (\Sigma_1 |v_i^2|)^{\frac{1}{2}}$ denote the length of the n-tuple $v = (v_1, \ldots, v_n)$.

If $n = 1$, then $W = [1]$ will satisfy (3).

If $n \geq 2$, let y be an n-tuple of unit length such that

$$\alpha_1(AB) = |ABy| = |By| \cdot \left| A\left(\frac{By}{|By|}\right)\right| . \tag{4}$$

But $|By| \leq \alpha_1(B)$ and $\left| A\left(\frac{By}{|By|}\right)\right| \leq \alpha_1(A)$. By hypothesis, $\alpha_1(AB) = \alpha_1(A)\alpha_1(B)$, so $|By| = \alpha_1(B)$. Since B is positive-definite hermitian. $By = \alpha_1(B)y$. Also $Ay = \alpha_1(A)y$.

Let S be a unitary matrix whose first column is y. Since S^*AS is hermitian, $S^*AS = \alpha_1(A) + A'$, where A' is an $(n-1)$-square, positive-definite hermitian matrix. Similarly, $S^*BS = \alpha_1(A) + B'$ where B' is an $(n-1)$-square positive-definite hermitian matrix. Clearly,

$$\alpha_i(A') = \alpha_{i+1}(A) \qquad 1 \leq i \leq n-1,$$

$$\alpha_i(B') = \alpha_{i+1}(B) \qquad 1 \leq i \leq n-1,$$

and

$$\alpha_i(A'B') = \alpha_{i+1}(AB) \qquad 1 \leq i \leq n-1.$$

So equality (1) implies that

$$\alpha_i(A'B') = \alpha_i(A')\alpha_i(B') \qquad 1 \leq i \leq k-1.$$

By the induction hypothesis applied to the $(n-1)$-square matrices A' and B', there exists an $(n-1)$-square unitary matrix S' such that

$$S'^*A'S' = \operatorname{diag}(\alpha_2(A), \ldots, \alpha_k(A)) \dot{+} T_1$$

and

$$S'^*B'S' = \operatorname{diag}(\alpha_2(A), \ldots, \alpha_k(A)) \dot{+} T_2$$

where T_1 and T_2 are $(n-k)$-square matrices.

Finally let $W = S(1 \dot{+} S')$, then

$$\begin{aligned} W^*AW &= (1 \dot{+} S'^*)S^*AS(1 \dot{+} S') \\ &= (1 \dot{+} S'^*)(\alpha_1(A) \dot{+} A')(1 \dot{+} S') \\ &= \alpha_1(A) \dot{+} S'^*A'S' \\ &= \operatorname{diag}(\alpha_1(A), \ldots, \alpha_k(A)) \dot{+} T_1. \end{aligned}$$

Similary $W^*BW = \operatorname{diag}(\alpha_1(B), \ldots, \alpha_k(B)) \dot{+} T_2$. *Q.E.D.*

We need the following definition in order to state Ostrowski's inequality.

DEFINITION: *Let $\phi(x_1, \ldots, x_k)$ be a real valued function of k real variables. ϕ is* ***convex*** *in a region R if*

$$\phi(\theta x + (1-\theta)y) \leq \theta\phi(x) + (1-\theta)\phi(y) \tag{5}$$

whenever $0 < \theta < 1$, $x = (x_1, \ldots, x_k) \in R$, $y = (y_1, \ldots, y_k) \in R$. If equality holds in (5) only when $x_i = y_i$, $1 \leq i \leq k$, we say ϕ is ***strictly convex****.*

Now we state two theorems by Ostrowski [3, Thm. XVI].

THEOREM 2 (Ostrowski): *Let* $f(x_1, \ldots, x_k)$ *be a symmetric function of* k *real variables such that*

$$\phi(x_1, \ldots, x_k) = f(\exp x_1, \ldots, \exp x_k)$$

is increasing in each variable x_i *and convex in the region* $x_i \geq 0$. *Then,*

$$f(\alpha_1(AB), \ldots, \alpha_k(AB)) \leq f(\alpha_1(A)\alpha_1(B), \ldots, \alpha_k(A)\alpha_k(B)). \tag{6}$$

THEOREM 3 (Ostrowski): *Suppose* $\phi(x_1, \ldots, x_k)$ *is a symmetric function which is convex and increasing in each variable. Let* $\{x_i, y_i\}$ *be* 2n *positive numbers satisfying*

$$\begin{gathered} x_1 \geq \ldots \geq x_n, \qquad y_1 \geq \ldots \geq y_n \\ x_1 + \ldots + x_r \leq y_1 + \ldots + y_r; \qquad 1 \leq r \leq n, \end{gathered} \tag{7}$$

with equality in (7) *for* r = n. *Then*

$$\phi(x_1, \ldots, x_k) \leq \phi(y_1, \ldots, y_k). \tag{8}$$

In addition, if ϕ *is strictly convex and strictly increasing in each variable, then equality holds in* (8) *if and only if* $x_i = y_i$, $1 \leq i \leq k$.

If the inequalities in (7) hold, x is said to *majorize* y, [1, pg. 45]. A proof of Theorem 3 different than Ostrowski's is given in [2]. Theorem 2 follows from Theorem 3 by choosing $x_i = \log \alpha_i(AB)$ and $y_i = \log(\alpha_i(A)\alpha_i(B))$. It is known that

$$\prod_{i=1}^{r} \alpha_i(AB) \leq \prod_{i=1}^{r} \alpha_i(A)\alpha_i(B), \qquad 1 \leq r \leq n.$$

Now using Theorem 1, it is easy to show that if ϕ is strictly convex and strictly increasing in each variable, then equality holds in Theorem 2, (6) if and only if there exists a unitary matrix W such that (2) holds.

References

[1] Hardy, G. H., Littlewood, J. E., and Pólya, G., Inequalities (second edition, Cambridge, 1952).

[2] Marcus, Marvin, and Gordon, William R., Analysis of equality in certain matrix inequalities (submitted).

[3] Ostrowski, A., Sur quelques applications des fonctions convexes et concaves au sens de I. Schur, Mathématiques Pures et Appliquees, **IV**, 253–292 (1952).

(Paper 74B4–337)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

Contractifiable Semigroups

Philip R. Meyers*

**Institute for Basic Standards, National Bureau of Standards,
Washington, D.C. 20234**

(September 17, 1970)

Consider a continuous semigroup of operators on a metric space, indexed $\{T_t\}$ by the nonnegative real numbers. It is shown that if any one of the operators can be made into a contraction by some topology-preserving remetrization, then for each $\lambda\epsilon(0, 1)$ there is a metric under which each operator $T_t(t > 0)$ becomes a contraction with contraction constant λ^t. With the operators regarded as describing the evolution of an autonomous dynamical system, this metric can be used to define a Lyapanov function.

Key words: Contractions; functional analysis; operators; semigroups; stability theory.

1. Introduction

This paper is the fifth in a series [1–4],[1] dealing with the question of when continuous self-mappings, of a metrizable topological space X, can be made into contractions by choosing a suitable metric for X. A single map f with this property will be termed *contractifiable*. It was shown in [2] that such maps f are characterized by the properties one would expect in view of Banach's Contraction Principle, namely the existence in X of a fixed point ξ of f and an open neighborhood U of ξ such that[2]

$$f^n(x) \to \xi \qquad (\text{all } x\epsilon X), \tag{1.1}$$

$$f^n(U) \to \{\xi\}. \tag{1.2}$$

Now let $\mathcal{T}$ be a commutative *family* of self-mappings of X, with common fixed-point ξ. Assume the members of $\mathcal{T}$ are individually contractifiable, so that there are collections $\{d_f : f\epsilon\mathcal{T}\}$ and $\{\lambda_f : f\epsilon\mathcal{T}\}$ such that each $f\epsilon\mathcal{T}$ is a contraction mapping of metric space (X, d_f) with $\lambda_f\epsilon(0, 1)$ as contraction constant. If d_f can be chosen independent of f, we term $\mathcal{T}$ *simultaneously contractifiable*. In [3] it is shown that $\mathcal{T}$ will indeed be simultaneously contractifiable, if it is *finite*. The present paper provides the analogous result for the most important special case of infinite families $\mathcal{T}$.

The special case is given by $\mathcal{T} = \mathcal{S} - \{T_0\}$, where $\mathcal{S} = \{T_t : t\epsilon R^+\}$ is an *R^+-semigroup of operators* on metric space (X, d). Here R^+ is the topological semigroup of nonnegative real numbers under addition, d is a metric on X (the "original" metric, prior to remetrization), and each T_t is a continuous self-mapping of X. The terminology means that T_0 is the identity map of X, that the semigroup property $T_{t+s} = T_t T_s$ is satisfied, and that the continuity property

$$\sup\{d(T_t(x), T_s(x)) : x\epsilon X\} \to 0 \qquad (\text{as } s \to t) \tag{1.3}$$

holds for all $t\epsilon R^+$. Our main result is the following.

AMS Subject Classification. Primary 20M99.

*Present address: Dynamic Access Corporation, 220 Park Avenue South, New York City, New York 10003.

[1] Figures in brackets indicate the literature references at the end of the paper.

[2] To avoid inessential technical complications we assume throughout that ξ actually lies in X, instead of being an "ideal point" corresponding to a class of equivalent Cauchy sequences.

THEOREM: *Let* $\mathcal{S}=\{T_t:t\epsilon R^+\}$ *be an* R^+*-semigroup of operators on* (X, d). *If any member* T_τ *of* $\mathcal{S}$ *is contractifiable, then for each* $\lambda\epsilon(0, 1)$ *there is a topologically equivalent metric* ρ_λ *such that each* $T_t(t>0)$ *is a contraction on* (X, ρ_λ) with λ^t *as contraction constant.*

This result must be distinguished from that in [4], where it was shown that there existed metrics $\rho_\lambda^{(t)}$ such that each $T_t(t>0)$ is a contraction on $(X, \rho_\lambda^{(t)})$ with λ as contraction constant. In particular, although the relationship of our topic with asymptotic stability was worked out in Theorem 2 of [4], it is only the above theorem which makes evident the association with a Lyapanov function for the semigroup $\mathcal{S}$ and the point $\xi\epsilon X$.

In this stability-theoretic context, X is interpreted as the space of possible "states" of an autonomous dynamical system, while $T_t(x)$ is interpreted as the system's state at time t if its initial state was x. The fixed-point property of ξ characterizes it as an "equilibrium state," whose stability is to be studied. In terms of the metric ρ_λ of the above Theorem, set

$$L(x)=\rho_\lambda(\xi, x).$$

Then L is continuous, and positive-definite with respect to ξ. Since

$$L(T_{s+t}(x))\leqslant\lambda^s L(T_t(x)),$$

L is nonincreasing along every trajectory, and strictly decreasing unless and until the trajectory reaches ξ. These properties identify L as a Lyapanov function.

2. Two Lemmas

The analysis will employ the notation

$$S(x, r)=\{y\epsilon X:d(x, y)\leqslant r\}. \tag{2.1}$$

It is convenient to precede the main proof by the following two lemmas, the first of which is a slight generalization of the lemma in [4].

LEMMA 1: *Given* $x\epsilon X$, $t>0$, *and* $\eta>0$, *there exists* $\delta>0$ *such that*

$$T_s(S(x, \delta))\subseteq S(T_s(x), \eta) \qquad \textit{for all } s\epsilon[0, t].$$

PROOF: If not, then there exist sequences $\delta_n\to 0$, $s(n)\epsilon[0, t]$, and $x_n\epsilon S(x, \delta_n)$ such that

$$d(T_{s(n)}(x), T_{s(n)}(x_n))>\eta. \tag{2.2}$$

By passing to a subsequence, we may assume $s(n)\to s\epsilon[0, t]$. Then

$$d(T_{s(n)}(x), T_{s(n)}(x_n))\leqslant d(T_{s(n)}(x), T_s(x))+d(T_s(x), T_s(x_n))+d(T_s(x_n), T_{s(n)}(x_n)).$$

By the continuity condition (1.3) and the fact $s(n)\to s$, the first and third summands on the right tend to 0 as $n\to\infty$; since T_s is continuous and $\delta_n\to 0$ implies $x_n\to x$, the same is true of the second summand. Thus the last display leads to a contradiction of (2.2).

LEMMA 2: *If some* T_τ *is contractifiable, then there is an open neighborhood* U *of* ξ *such that*

$$T_t(U)\subseteq U \qquad \textit{for all } t\geqslant 0, \tag{2.3}$$

$$T_t(U)\to\{\xi\} \qquad \textit{as } t\to\infty. \tag{2.4}$$

PROOF: Since T_τ is contractifiable, there is an open neighborhood U_τ of ξ such that $T_\tau^n(U_\tau) \to \{\xi\}$ as $n \to \infty$. As shown in [2], it can further be assumed that $T_\tau(U_\tau) \subseteq U_\tau$. For each $s \in [0, \tau]$, define the open set

$$U_s^* = \{x \in X : T_s(x) \in U_\tau\}$$

and put

$$U = \cap \{U_s^* : s \in [0, \tau]\}.$$

To prove (2.3) for any given $t \geq 0$, choose any $s \in [0, \tau]$ and write

$$t + s = n\tau + \sigma \qquad (\sigma \in [0, \tau];\ \text{integral } n \geq 0).$$

Then since $U \subset U_\sigma^*$, we have

$$T_{t+s}(U) = T_\tau^n T_\sigma(U) \subseteq T_\tau^n(U_\tau) \subseteq U_\tau,$$

so that $T_t(U) \subseteq U_s^*$. Hence (2.3) holds.

To prove (2.4), consider any $\eta > 0$. By Lemma 1, there is a $\delta > 0$ such that

$$T_s(S(\xi, \delta)) \subseteq S(\xi, \eta) \qquad \text{for all } s \in [0, \tau].$$

Choose N so large that

$$T_\tau^n(U_\tau) \subseteq S(\xi, \delta) \qquad \text{for all } n \geq N.$$

Then for $t \geq t_0(\eta) = N\tau$ we have

$$t = n\tau + \sigma \qquad (n \geq N;\ \sigma \in [0, \tau)),$$

and so

$$T_t(U) \subseteq T_t(U_0^*) = T_t(U_\tau) = T_\sigma T_\tau^n(U_\tau)$$

$$\subseteq T_\sigma(S(\xi, \delta)) \subseteq S(\xi, \eta).$$

This establishes (2.4).

Clearly $\xi \in U$; to prove U an open neighborhood of ξ, it remains to show that U is open. Consider any $x \in U$. We assert the existence of a $\delta > 0$ such that

$$T_s(S(x, \delta)) \subseteq U_\tau \qquad \text{for all } s \in [0, \tau]; \tag{2.5}$$

this implies $S(x, \delta) \subseteq U$, and thus that U is open. If no such δ existed, there would be sequences $\delta_n \to 0$, $s(n) \in [0, \tau]$ and $x_n \in S(x, \delta_n)$ such that $T_{s(n)}(x_n) \in X - U_\tau$. By passing to a subsequence, we may assume $s(n) \to s \in [0, \tau]$. Then

$$d(T_{s(n)}(x_n), T_s(x)) \leq d(T_{s(n)}(x_n), T_s(x_n)) + d(T_s(x_n), T_s(x));$$

arguing as in the proof of Lemma 1 we find that $T_{s(n)}(x_n) \to T_s(x)$. This however contradicts the fact that each $T_{s(n)}(x_n)$ lies in the *closed* set $X - U_\tau$, whereas $x \in U$ implies $T_s(x) \notin U_\tau$.

3. Proof of Theorem

The proof follows that in [2], for which the underlying semigroup was that of the nonnegative integers rather than R^+, but the additional details which are needed warrant a full account. The first step is to construct a metric ρ_m, topologically equivalent to d, with respect to which each T_t is *nonexpanding* in the sense that

$$\rho_m(T_t(x), T_t(y)) \leq \rho_m(x, y) \qquad (x, y \in X). \tag{3.1}$$

This is accomplished by setting

$$\rho_m(x, y) = \sup \{d(T_t(x), T_t(y)) : t \geq 0\}. \tag{3.2}$$

It will initially be shown that (3.2) yields a finite result. For each $x \epsilon X$, observe that

$$C_x = \{T_\sigma(x) : 0 \leq \sigma \leq \tau\}$$

is a continuous image of $[0, \tau]$, and hence is compact. Let ρ be a metric on X such that T_τ is a contraction on (X, ρ), with $\theta \epsilon (0, 1)$ as contraction constant. Then

$$M_x = \max \{\rho(\xi, z) : z \epsilon C_x\}$$

is finite. Writing any $t \geq 0$ as

$$t = n\tau + \sigma \qquad (\sigma \epsilon [0, \tau]; \text{ integral } n \geq 0), \tag{3.3}$$

we obtain

$$\rho(T_t(x), \xi) = \rho(T_\tau^n T_\sigma(x), T_\tau^n(\xi)) \leq \theta^n \rho(T_\sigma(x), \xi) \leq \theta^n M_x.$$

Since $n \to \infty$ as $t \to \infty$, it follows that

$$T_t(x) \to \xi \qquad \text{as } t \to \infty. \tag{3.4}$$

For particular $x, y \epsilon X$, this implies that

$$d(T_t(x), T_t(y)) \leq d(T_t(x), \xi) + d(\xi, T_t(y)) \to 0$$

as $t \to \infty$. Thus there is a $t_0 > 0$ such that

$$d(T_t(x), T_t(y)) < 1 \qquad \text{for all } t > t_0.$$

Since $d(u, v)$ must be bounded on the compact subset

$$\{(T_t(x), T_t(y)) : 0 \leq t \leq t_0\}$$

of $X \times X$, it follows that (3.2) indeed yields a finite result.

The positive definiteness and symmetry of ρ_m follow directly from the corresponding properties of d. The triangle inequality for ρ_m follows from the fact that, for any $t \geq 0$,

$$d(T_t(x), T_t(y)) \leq d(T_t(x), T_t(z)) + d(T_t(z), T_t(y)) \leq \rho_m(x, z) + \rho_m(z, y).$$

Thus ρ_m is indeed a metric on the set carrying space X. That ρ_m has the nonexpanding property (3.1) follows easily from (3.2). Thus it only remains to prove that ρ_m and d are topologically equivalent. Since $\rho_m \geq d$, we have

$$\rho_m(x_n, x) \to 0 \qquad \text{implies} \qquad d(x_n, x) \to 0,$$

and so it suffices to prove the converse implication.

Consider then any $x \epsilon X$ and any $\eta > 0$. With U as in Lemma 2, there is no loss of generality for what follows in assuming η small enough that

$$S(\xi, \eta) \subseteq U. \tag{3.5}$$

By (3.4), there is a $t(0) > 0$ such that

$$d(\xi, T_t(x)) < \eta/2 \quad \text{for all} \quad t \geq t(0), \tag{3.6}$$

and by Lemma 1 there is a $\delta' > 0$ such that

$$T_s(S(x, \delta')) \subseteq S(T_s(x), \eta/2) \quad \text{for all} \quad s \in [0, t(0)]. \tag{3.7}$$

Since $T_t(U) \to \{\xi\}$ as $t \to \infty$, there is a $t(1) > 0$ such that

$$T_t(U) \subseteq S(\xi, \eta/2) \qquad \text{for all } t > t(1); \tag{3.8}$$

and by Lemma 1 there is a $\delta'' > 0$ such that

$$T_s(S(x, \delta'')) \subseteq S(T_s(x), \eta/2) \qquad \text{for all } s \in [0, t(0) + t(1)]. \tag{3.9}$$

Let $\delta = \min(\delta', \delta'')$; we will show that

$$T_t(S(x, \delta)) \subseteq S(T_t(x), \eta) \qquad (\text{all } t \geq 0), \tag{3.10}$$

and establishing the existence (for any x and η) of a δ with this last property completes the current step of the proof.

That (3.10) holds for $0 \leq t \leq t(0) + t(1)$ follows from (3.9). Suppose then that $t > t(0) + t(1)$. By (3.7) with $s = t(0)$, (3.6) with $t = t(0)$, and (3.5), we have

$$T_{t(0)}(S(x, \delta)) \subseteq S(\xi, \eta) \subseteq U,$$

and since $t - t(0) > t(1)$, (3.8) yields

$$T_t(S(x, \delta)) = T_{t-t(0)} T_{t(0)}(S(x, \delta)) \subseteq T_{t-t(0)}(U) \subseteq S(\xi, \eta/2),$$

implying (3.10).

For the *second* step in constructing the desired metric ρ_λ, we begin by observing that the properties (2.3) and (2.4) of U imply

$$T_t(\bar{U}) \subseteq \bar{U}; \qquad \overline{T_t(U)} \to \{\xi\} \qquad \text{as } t \to \infty.$$

We now define

$$\nu(x) = \sup \{t : x \in \overline{T_t(U)}\} \qquad \text{for } x \in \bar{U},$$

so that $\nu(\xi) = \infty$ and $0 \leq \nu(x) < \infty$ for $x \in \bar{U} - \{\xi\}$, and also set

$$\nu(x) = -\inf \{t : x \in T_t^{-1}(\bar{U})\} \qquad \text{for } x \in X - \bar{U},$$

so that $(-\infty) < \nu(x) < 0$ in this case.[3] Using the continuity of the semigroup, it is readily shown that the extrema in these definitions of $\nu(x)$ are actually attained. For our purposes, the critical property of $\nu(x)$ is

$$\nu(T_t(x)) \geq \nu(x) + t, \tag{3.11}$$

[3] The finiteness of $\nu(x)$ is an easy consequence of (3.4).

with verification straightforward.[4] With the further definition

$$c(x, y) = \min \{\nu(x), \nu(y)\} \tag{3.12}$$

it follows from (3.11) that

$$c(T_t(x), T_t(y)) \geq c(x, y) + t. \tag{3.13}$$

Now define

$$d_\lambda(x, y) = \lambda^{c(x, y)} \rho_m(x, y), \tag{3.14}$$

which has the correct limiting form $d_\lambda(\xi, \xi) = 0$. Then d_λ is positive definite and symmetric, and by (3.13) also has the property

$$d_\lambda(T_t(x), T_t(y)) \leq \lambda^t d_\lambda(x, y) \tag{3.15}$$

desired for ρ_λ. However, d_λ may not satisfy the triangle inequality.

This is rectified in the *third* and last step of the construction. Denote by Σ_{xy} the set of chains

$$\sigma_{xy} = [x = x_0, x_1, \ldots, x_m = y]$$

from x to y, with associated *lengths*

$$L_\lambda(\sigma_{xy}) = \sum_1^m d_\lambda(x_i, x_{i-1}),$$

and put

$$\rho_\lambda(x, y) = \inf \left\{ L_\lambda(\sigma_{xy}) : \sigma_{xy} \epsilon \Sigma_{xy} \right\}.$$

We shall show that ρ_λ is the desired metric.

The contraction property

$$\rho_\lambda(T_t(x), T_t(y)) \leq \lambda^t \rho_\lambda(x, y)$$

follows by applying (3.15) to the links $[x_{i-1}, x_i]$ of any chain σ_{xy}. Clearly ρ_λ is symmetric and $\rho_\lambda(x, x) = 0$; the triangle law holds since following a σ_{xy} with a σ_{yz} yields a σ_{xz}.

It remains to show that ρ_λ is positive definite. For $(-\infty) < \nu \leq \infty$, define the closed set

$$K_\nu = \{x \epsilon X : \nu(x) \geq \nu\}; \tag{3.16}$$

the sets $\{K_\nu\}$ are a nonascending family, with limit $\{\xi\}$.

Consider any $x \neq \xi$ and any $y \neq x$; assume $\nu(x) \leq \nu(y)$ without loss of generality. Choose any $s > 0$. First suppose $y \neq \xi$; then $y \epsilon X - K_{\nu(y)+s}$. Thus any chain σ_{xy} either has all its points in $X - K_{\nu(y)+s}$, implying

$$L_\lambda(\sigma_{xy}) \geq \lambda^{\nu(y)+s} \rho_m(x, y),$$

or else has a last among its links which leaves $K_{\nu(y)+s}$ (and possibly is followed by other links), implying

$$L_\lambda(\sigma_{xy}) \geq \lambda^{\nu(y)+s} \rho_m(y, K_{\nu(y)+s}).$$

It follows that

$$\rho_\lambda(x, y) \geq \lambda^{\nu(y)+s} \min \{\rho_m(x, y), \rho_m(y, K_{\nu(y)+s})\} > 0. \tag{3.17}$$

---

[4] E.g., for $\nu(x) \geq 0$ we have $x \epsilon \overline{T_{\nu(x)}(U)}$, implying $T_t(x) \epsilon T_t(\overline{T_{\nu(x)}(U)})$ and thus $T_t(x) \epsilon \overline{T_t T_{\nu(x)}(U)} = \overline{T_{\nu(x)+t}(U)}$, from which (3.11) follows.

The remaining case, $y=\xi$, is covered by applying analogous reasoning to yield

$$\rho_\lambda(x, \xi) \geq \lambda^{\nu(x)+s}\rho_m(x, K_{\nu(x)+s}) > 0. \tag{3.18}$$

By (3.17) and (3.18), ρ_λ is positive definite and hence indeed a metric, which must still be proved equivalent to ρ_m.

Now define

$$\mu(x) = \inf\{\mu \geq 0 : T_\mu(x) \epsilon U\} < \infty; \tag{3.19}$$

it follows from the definition of $\nu(x)$ that

$$\nu(x) \geq -\mu(x). \tag{3.20}$$

With the further definition

$$B_\mu = X - T_\mu^{-1}(U) \tag{3.21}$$

we have $\rho_m(x, B_{\mu(x)+s}) > 0$ for all $s > 0$.

Consider any $x \epsilon X$; we will show that

$$\rho_m(x_n, x) \to 0 \quad \text{implies} \quad \rho_\lambda(x_n, x) \to 0. \tag{3.22}$$

First suppose $x \neq \xi$; choose any $s > 0$. Consider any $y \epsilon X$ with

$$\rho_m(x, y) < \rho_m(x, B_{\mu(x)+s}). \tag{3.23}$$

Then $y \epsilon X - B_{\mu(x)+s}$, implying $T_{\mu(x)+s}(y) \epsilon U$ and hence that $\mu(x) + s \geq \mu(y)$. By (3.20), this implies $\nu(y) \geq -\mu(x) - s$. Since (3.20) also implies $\nu(x) \geq -\mu(x) - s$, we have $c(x, y) \geq -\mu(x) - s$, so that (3.23) implies the second inequality in

$$\rho_\lambda(x, y) \leq d_\lambda(x, y) \leq \lambda^{-\mu(x)-s}\rho_m(x, y). \tag{3.24}$$

Since (3.23) implies the indicated inequality between the first and third terms of (3.24), (3.22) has been established for all $x \neq \xi$. As for the case $x = \xi$, note that if $\rho_m(y, \xi) < \rho_m(\xi, B_0)$, then $\nu(y) \geq 0$, and so $c(\xi, y) \geq 0$ and

$$\rho_\lambda(\xi, y) \leq d_\lambda(\xi, y) \leq \rho_m(\xi, y), \tag{3.25}$$

showing that (3.22) holds in this case also.

It only remains to prove that

$$\rho_\lambda(x_n, x) \to 0 \quad \text{implies} \quad \rho_m(x_n, x) \to 0. \tag{3.26}$$

As above, first assume $x \neq \xi$, and choose any $s > 0$. Since $K_\nu \to \{\xi\}$, the argument culminating in (3.25) shows that there exists $k(x) > \max\{0, \nu(x) + s\}$ such that $z \epsilon K_{k(x)}$ implies $\rho_\lambda(\xi, z) < \rho_\lambda(x, \xi)/2$. Then

$$\rho_\lambda(x, K_{k(x)}) \geq \rho_\lambda(x, \xi)/2. \tag{3.27}$$

Consider any $y \epsilon X$ for which

$$\rho_\lambda(x, y) < \rho_\lambda(x, \xi)/2. \tag{3.28}$$

It follows from (3.27) that in forming $\rho_\lambda(x, y)$ as an infimum, only chains disjoint from $K_{k(x)}$ need be considered, implying

$$\rho_\lambda(x, y) \geq \lambda^{k(x)}\rho_m(x, y). \tag{3.29}$$

Now suppose that y, besides obeying (3.28), also satisfies

$$\rho_\lambda(x, y) < \lambda^{k(x)}\rho_m(x, K_{\nu(x)+s}). \tag{3.30}$$

Then (3.29–30) imply

$$\rho_m(x, y) < \rho_m(x, K_{\nu(x)+s}). \tag{3.31}$$

We will show that

$$\rho_m(x, y) \leq \lambda^{-\nu(x)-s}\rho_\lambda(x, y); \tag{3.32}$$

that this is implied by (3.28) and (3.30), shows that (3.26) indeed holds for $x \neq \xi$.

In case $\nu(y) \leq \nu(x)$, (3.32) is obtained by applying (3.17) with x and y reversed, and then taking account of (3.31). In the contrary case, (3.31) implies $y \epsilon X - K_{\nu(x)+s}$; thus any chain σ_{xy} either lies in $X - K_{\nu(x)+s}$, implying

$$L_\lambda(\sigma_{xy}) \geq \lambda^{\nu(x)+s}\rho_m(x, y),$$

or else has a first among its links which enters $K_{\nu(x)+s}$, implying

$$L_\lambda(\sigma_{xy}) \geq \lambda^{\nu(x)+s}\rho_m(x, K_{\nu(x)+s}).$$

Taking (3.31) into account, it follows that

$$\rho_\lambda(x, y) \geq \lambda^{\nu(x)+s}\rho_m(x, y),$$

which is equivalent to the desired result (3.32).

Finally, (3.26) must be verified in the case $x = \xi$. Note that for any $\eta > 0$, there is a $\nu > 0$ such that $\rho_m(\xi, z) < \eta/2$ for all $z \epsilon K_\nu$. Consider any $y \epsilon X$ such that

$$\rho_m(\xi, y) > \eta; \tag{3.33}$$

it follows that $\nu(y) < \nu$ and $\rho_m(y, K_\nu) \geq \eta/2$. Choose $s > 0$ so that $\nu(y) + s = \nu$; then (3.18) yields the first inequality in

$$\begin{aligned}\rho_\lambda(\xi, y) &\geq \lambda^\nu \rho_m(y, K_{\nu(y)+s}) \\ &= \lambda^\nu \rho_m(y, K_\nu) \geq \lambda^\nu \eta/2 \equiv \delta > 0.\end{aligned} \tag{3.34}$$

The contrapositive, of the implication of (3.34) by (3.33), is that $\rho_\lambda(\xi, y) < \delta$ implies $\rho_m(\xi, y) \leq \eta$. Since for each $\eta > 0$ there is a $\delta > 0$ with this property, (3.26) has indeed been verified for $x = \xi$.

4. References

[1] Meyers, P. R., Some extensions of Banach's Contraction Theorem, J. Res. Nat. Bur. Stand. (U.S.), **69B** (Math. Phys.), No. 3, 179–184 (July–Sept. 1965).

[2] Meyers, P. R., A converse to Banach's Contraction Theorem, J. Res. Nat. Bur. Stand. (U.S.), **71B** (Math. Phys.), Nos. 2 & 3, 73–76 (April–Sept. 1967).

[3] Goldman, A. J., and Meyers, P. R., Simultaneous contractification, J. Res. Nat. Bur. Stand. (U.S.), **73B** (Math. Sci.), No. 4, 301–305 (Oct.–Dec. 1969).

[4] Meyers, P. R., On contractive semigroups and uniform asymptotic stability, J. Res. Nat. Bur. Stand. (U.S.), **74B** (Math. Sci.), No. 2, 115–120 (Apr.–June 1970).

(Paper 74B4–338)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

Publications of the National Bureau of Standards*

Selected Abstracts

Olver, F. W. J., **Why steepest descents?**, *SIAM Rev.* **12**, *No. 2, 228–247 (April 19, 1970).*

Key words: Airy integral; asymptotics; error bounds; gamma function; saddle-points; steepest descents; Watson's lemma.

A theorem is proved concerning the asymptotic expansion of contour integrals of the form

$$\int_a^b e^{-zp(t)}q(t)dt$$

in which $p(t)$ and $q(t)$ are analytic functions of t, and z is a large complex parameter. In comparison with the method of steepest descents, the essential condition imposed on the integration path by the new theorem is simply that $\mathrm{Re}\{zp(t)\}$ attains its minimum at one of the endpoints. The theorem is illustrated by application to the Airy function of large complex argument.
Methods are given for the calculation of realistic bounds for the error terms in the asymptotic expansion. In this analysis the use of paths of steepest descent is necessary, at least in the neighborhood of the minimum of $\mathrm{Re}\{zp(t)\}$. The error analysis is illustrated by some new results concerning Stirling's series.

Osgood, C. F., **A method in diophantine approximation (III)**, *Acta Arithmetica* **XVI**, 5–22 (1969).

Key words: Algebraic numbers; convolution product; differential equation; diophantine approximation.

In this paper the class of functions discussed in Part II is investigated further. After introducing several normalizing conditions, it is shown to be a module under convolution multiplication from a certain ring of functions. Also algebraic coefficients and algebraic points can now be handled in our original type of differential equation (see (1) on pp. 1).

Osgood, C. F., **A method in diophantine approximation (IV)**, Acta Arithmetica **XVI**, 23–40 (1969).

Key words: Convolution product; differential equations; diophantine approximation; irrational numbers; n-complex variables; rational numbers.

This paper generalizes the results in part III of this series of papers to $n > 1$ complex variables. It is shown that the values of a certain class of functions of n variables are poorly approximable at rational points in an appropriate domain. The algebraic structure of this class of functions is studied in an effort to determine the composition of the class. A module structure is exhibited under convolution product, i.e.,

$$g_1(z_1, \ldots, z_n) {}_k^* g_2(z_1, \ldots, z_n)$$

$$= \int_0^{z_k} g_1(z_1, \ldots, z_{k-1}, t_k, z_{k+1}, \ldots, z_n)$$

$$g_2(z_1, \ldots, z_{k-1}, z_k - t_k, z_{k+1}, \ldots, z_n)\partial t.$$

Other NBS Publications

J. Res. Nat. Bur. Stand. (U.S.), 74A (Phys. and Chem.), No. 5 (September–October 1970), SD Catalog No. C13.22/sec. A:74/5.

Heat capacity and thermodynamic properties of α-beryllium nitride, Be_3N_2, from 20 to 315 K. G. T. Furukawa and M. L. Reilly.
Heat capacity and thermodynamic properties of lithium hexafluoroaluminate, Li_3AlF_6, from 15 to 380 K. G. T. Furukawa, W. G. Saba, and J. C. Ford.
Thermodynamics of hydrochloric acid in 80 weight percent 2-methoxyethanol and 20 weight percent water from 10 to 50 °C. H. P. Thun, B. R. Staples, and R. G. Bates.
Vapor pressure and heat of sublimation of platinum. E. R. Plante, A. B. Sessoms, and K. R. Fitch.
Thermophysical properties of methane: Virial coefficients, vapor- and melting pressures. R. D. Goodwin.
Specific heats, C_v, of compressed liquid and gaseous fluorine. Rolf Prydz and R. D. Goodwin.
Acidic dissociation of diprotonated piperazine in methanol-water solvents from 10 to 40 °C. M. Paabo and R. G. Bates.
Thermal conductivity standard reference materials from 4 to 300 K. I. Armco iron: Including apparatus description and error analysis. J. G. Hust, R. L. Powell, and D. H. Weitzel.
Mössbauer spectrometer calibration using TiFe. L. J. Swartzendruber and L. H. Bennett.
Energies of 1snl ($l \geq 3$) configurations in He^4 I as calculated from polarization theory. W. C. Martin.
Odd configurations in neutral nickel (Ni I). Charles Roth.
A theoretical investigation of the configurations $(3d+4s)^9 4p$ in neutral cobalt (Co I). Charles Roth.
Electron-spin resonance study of an alkaline solution of copper (II) oxalate—*meso*-tartrate complex (the Somogyi reagent) and related complexes. A. J. Fatiadi.

J. Res. Nat. Bur. Stand. (U.S.), 74A (Phys. and Chem.), No. 6 (November–December 1970), SD Catalog No. C13.22/sec. A:74/6.

PVT measurements, virial coefficients, and Joule-Thomson inversion curve of fluorine. R. Prydz and G. C. Straty.
The activity coefficients of hydrofluoric acid in water from 0 to 35 °C. W. J. Hamer and Y.-C. Wu.
Fluorine flame calorimetry III. The heat of formation of chlorine trifluoride at 298.15 K. R. C. King and G. T. Armstrong.
A review of oscillator strengths for lines of Cu I. Charles H. Corliss.
Stark effect and hyperfine structure of HCN measured with an electric resonance maser spectrometer. H. E. Radford and C. V. Kurtz.
The second spectrum of nickel (ni II). A. G. Shenstone.

J. Res. Nat. Bur. Stand. (U.S.), 74C (Eng. & Instr.), Nos. 1 & 2 (January–June 1970), SD Catalog No. C13.22/sec. C:74/1 & 2.

Production of near-perfect interferograms of variable visibility. J. B. Saunders.
Locked nuclear quadrupole resonance spectrometer for pressure measurements. R. C. Frisch and D. L. VanderHart.
Precise continuous optical attenuator. G. Ruffino.
An A–C resistance thermometer bridge. Robert D. Cutkosky.
On the flow induced by a Maxwell-Chartoff rheometer. E. A. Kearsley.
Mechanical compliance measurements of single-edge-notch tension specimens. M. J. Orloski.
Effect of notch geometry and temperature on the creep-rupture behavior of a titanium alloy. W. D. Jenkins and W. A. Willard.

J. Res. Nat. Bur. Stand. (U.S.), 74C (Eng. & Instr.), Nos. 3 & 4 (July-December 1970), SD Catalog No. C13.22/sec. C:74/3 & 4.

Techniques for comparing four-terminal-pair admittance standards. R. D. Cutkosky.

Some applications for series impedance elements in radio frequency immittance measurements. L. E. Huntley.

A pulse heating method for the measurement of melting point of electrical conductors (thin wires) above 2000 K. A. Cezairliyan.

A time-shared computer system for diffractometer control. H. A. Alperin and E. Prince.

Tensile deformation of vapor-deposited copper reinforced with tungsten wires. W. D. Jenkins, W. A. Willard, and D. A. Harne.

"Connector-pair" techniques for the accurate measurement of two-terminal low-value capacitances. A. Millea.

The use of dew-point temperature in humidity calculations. L. A. Wood.

Bolovac application for HF and microwave power measurement and standardization. M. C. Selby.

Alley, C. O., Chang, R. F., Currie, D. G., Poultney, S. K., Bender, P. L., Dicke, R. H., Wilkinson, D. T., Faller, J. E., Kaula, W. M., MacDonald, J. F., Mulholland, J. D., Plotkin, H. H., Carrion, W., Wampler, E. J., Laser ranging retro-reflector: Continuing measurements and expected results, Science **167,** No. 3918, 458–460 (Jan. 30, 1970).

American National Standard, Radiological safety in the design and operation of particle accelerators, NBS Handb. 107, 22 pages (June 1970) 30 cents, SD Catalog No. C13.11:107.

Anderson, R. L., Guildner, L. A. Edsinger, R. E., Movable and fixed modular vacuum devices with confined fluorocarbon plastic seals, Rev. Sci. Instr. **41,** No. 7, 1076–1082 (July 1970).

Arenhovel, H., Danos, M., Williams, H. T., Baryon resonances in nuclei: magnetic moments of the deuteron, Physics Letters **31B,** No. 3, 109–112 (February 2, 1970).

Arp, V., Heat transport through helium II, Cryogenics **10,** No. 2, 96–105 (April 1970).

Astin, A. V., Report on the symposium on an International Standard Reference Materials Program, Metrologia **6,** No. 1, 33–34 (Jan. 1970).

Baker, M. A., A Summary of the three year inspection of the 1964 exposure test of porcelain enamels on aluminum, (Chaper in Proceedings of the Porcelain Enamel Institute Technical Forum, Ohio State Univ., Columbus, Ohio, Nov. 4–6, 1969, 31, 106–113 (Porcelain Enamel Institute, Washington, D.C., 1969).

Baker, M. A., Continuity of coating test and suggested applications, Chapter in Proceedings of the Procelain Enamel Institute Technical Forum, Ohio State Univ., Columbus, Ohio, No. 4–6, 1969, 31, 71–84 (Porcelain Enamel Institute, Washington, D.C. 1969).

Ballard, D. B., Yakowitz, H., Investigation of secondary cracks from the failed Point Pleasant, West Virginia Bridge, Proc. Third Annual Scanning Electron Microscope Symp., April 1970, 321–328 (Illinois Institute of Technology, Research Institute, Chicago, Ill., 1970).

Bennett, L. H., Swartzendruber, L. J., On the interpretation of Mössbauer effect spectra as related to the constitution of Cu-Ni-Fe alloys, Acta Met. **18,** No. 5, 485–498 (May 1970).

Benson, S. W., O'Neal, H. E., Kinetic data on gas phase unimolecular reactions, NBS NSRDS–21, 645 pages (Feb. 1970) $7.00, SD Catalog No. C13.48:21.

Berger, M. J., Seltzer, S. M., Maeda, K., Energy deposition of auroral electrons in the atmosphere, J. Atmospheric Terrest. Phys. **32,** 1015–1045 (1970).

Berger, M. J., Spectrum of energy deposited electrons in spherical regions, Proc. 2nd Symp. on Microdosimetry, Stresa, Italy, October 20–24, 1969, Euratom Document No. 4452 d-f-e, pp. 541–559 (1970).

Berry, V. M., Sage, B. H., Phase behavior in binary and multicomponent systems at elevated pressures: n-pentane and methane-n-pentane, Nat. Bur. Stand. (U.S.), Nat. Stand. Ref. Data. Ser. 32, 73 pages (June 1970) 70 cents, SD Catalog No. C13.48:32.

Bjorge, S. E. Urbach, P. F., Earl, P. H., King, D. W., Wiederkehr, R. R. V., An experiment to determine the effectiveness of various announcement media on Clearinghouse For Federal Scientific and Technical Information sales, Proc. American Society for Information Science Annual Meeting, Columbus, Ohio, Oct. 20–24, 1968, pp. 327–329 (Greenwood Publ. Corp., New York, N.Y., 1969).

Bjorge, S. E., Urbach, P. F., Earl, P. H., King, D. W., Wiederkehr, R. R. V., Experimentation, modeling and analysis to establish a new pricing policy at the Clearinghouse for Federal Scientific and Technical Information, Proc. American Society for Information Science Annual Meeting, Columbus, Ohio, Oct. 20–24, 1968, pp. 311–314 (Greenwood Publ. Corp., New York, N.Y., 1969).

Blaine, R. L., Tomes, L. A., Exploratory studies of early strength development in portland cement pastes and mortars, Nat. Bur. Stand. (U.S.), Bldg. Sci. Ser. 28, 14 pages (July 1970) 25 cents, SD Catalog No. C13.29/2:28.

Blake, R. W., Language of performance, an NBS project, Mater. Res. Std. **9,** No. 3, 11–14 (Mar. 1969).

Boesch, L., Napolitano, A., Macedo, P. B., Spectrum of volume relaxation times in B_2O_3, J. Am. Ceram. Soc. **53,** No. 3, 148–153 (1970).

Boswarva, I. M., Semiempirical calculations of ionic polarizabilities and van der Waals potential coefficients for the alkaline-earth chalcogenides, Phys. Rev. B. **1,** No. 4, 1698–1701 (Feb. 15, 1970).

Branscomb, L. M., The photoionization of one- and two-electron atoms (Proc. Sommerfeld Centennial Memorial Meeting, Munich, Germany, Sept. 10–14, 1968), Chapter in Physics of the One- and Two-Electron Atoms, F. Bopp and H. Kleinpoppen, Ed., pp. 669–699 (North Holland Publ. Co., Amsterdam, The Netherlands, 1969).

Brauer, G. M., Pyrolysis-gas chromatographic techniques for polymer identification, Chapter 2. Thermal Characteristics Techniques in Techniques and Methods of Polymer Evaluations, P. Slade, Ed., II, 41–104 (Marcel Dekker Publ., New York, N.Y., 1970).

Braun, W., Bass, A. M., Pilling, M., Flash photolysis of ketene and diazomethane: the production and reaction kinetics of triplet and singlet methylene, J. Chem. Phys. **52,** No. 10, 5131–5143 (May 15, 1970).

Brenner, A., Sligh, J. L., Electrodeless electrolysis, J. Electrochem. Soc. **117,** No. 5, 602–608 (May 1970).

Bridges, J. M., Wiese, W. L., The oscillator-strength scale for Fe I, Astrophys. J. **161,** L71–L75 (July 1970).

Broadhurst, M. G., Mopsik, F. I., Bulk modulus and Grüneisen parameters for linear polymers, J. Chem. Phys. **52,** No. 7, 3634–3641 (Apr. 1, 1970).

Brombacher, W. G., Survey of micromanometers, Nat. Bur. Stand. (U.S.), Monogr. 114, 62 pages (June 1970) 60 cents, SD Catalog No. C13.44:114.

Brower, W. S., Jr., Fang, P. H., Dielectric constants of scheelite structure crystals, J. Appl. Phys. **40,** No. 12, 4988–4989 (November 1969).

Brower, W. S., Jr., Fang, P. H., Dielectric constants of zinc tungstate, J. Appl. Phys. **41,** No. 5, 2266 (April 1970).

Brower, W. S., Jr., Farabaugh, E. N., Dislocation etchant for single crystal Y_2O_3, J. Am. Ceram. Soc. **53,** No. 4, 225 (April 1970).

Cezairliyan, A., High-speed methods of measuring specific heat of electrical conductors at high temperatures (A review), High Temperatures–High Pressures **1,** 517–529 (1969).

Bryson, J. O., Carpenter, E. F., Flexural behavior of prestressed concrete composite Tee-beams, Nat. Bur. Stand. (U.S.), Bldg. Sci. Ser. 31, 14 pages (July 1970) 25 cents, SD Catalog No. C13.29/2:31. Catalog of standard reference materials, Nat. Bur. Stand. (U.S.), Spec. Publ. 260–1970 Ed., 84 pages (July 1970) 75 cents, SD Catalog No. C13.10:260–1970.

Bullis, W. M., Editor, Methods of measurement for semiconductor materials, process control, and devices, Quarterly Report Oct. 1 to Dec. 31, 1969, NBS Tech. Note, 527, 60 pages (May 1970) 60 cents, SD Catalog No. C13.46:527.

Bullough, R., Simmons, J. A., On the deformation of an imperfect solid, Chapter 5 in Physics of Strength and Plasticity, A. S. Argon, Ed., 47–63 (Massachusetts Institute of Technology Press, Cambridge, Mass., 1969).

Bur, A. J., Extrapolation to single-relaxation-time behavior in solutions of poly(*n*-butylisocyanate), J. Chem. Phys. **52,** No. 7, 3813–3819 (Apr. 1, 1970).

Burgess, A., Hummer, D. G., Tully, J. A., Electron impact excitation of positive ions, Phil. Trans. Roy. Soc. London, Ser. **A.,** 266, No. 1175, 225–279 (Apr. 23, 1970).

Burke, P. G., Gallaher, D. F., Geltman, S., Electron scattering by atomic hydrogen using a pseudo-state expansion. I. Elastic scattering, J. Phys. B. (Proc. Phys. Soc.) **2**, 1142 (1969).

Burke, R. W., Exchange reactions of ternary ion association complexes directly in the organic phase, Talanta **17**, 240–242 (1970).

Burke, R. W., Deardorff, E. R., Simultaneous spectrophotometric determination of cobalt, nickel and copper and 2,3-quinoxaline-Dithiol, Talanta 17, 255–264 (1970).

Burns, G. W., Hurst, W. S., An investigation of W–3% Re and W–25% Re thermoelements in vacuum argon and hydrogen, NASA Final Report CR–72639, 1–28 (National Aeronautics and Space Administration, Greenbelt, Md., Mar. 10, 1970).

Cassidy, E. C., Cones, H. N., Electro-optical observations and measurements of distorted high-intensity electric fields, Proc. Conf. Electrical Insulation and Dielectric Phenomena, Buck Hill Falls, Pa., Oct. 20–22, 1969, pp. 77–86 (National Academy of Sciences, Washington, D.C., 1970).

Cassidy, E. C., Cones, H. N., Use of expanded laser beam to analyze high-intensity electric fields, J. Soc. Motion Picture Television Engrs. **79**, No. 7, 590–591 (July 1970).

Caswell, R. S., Danos, M., On the accuracy of the adiabatic separation method, J. Math. Phys. **11**, No. 2, 349–354 (Feb. 1970).

Chandler, R. F., Christian, R. A., Crash testing of humans in automobile seats, Proc. Society of Automotive Engineers Safety Conf. May 13–15, 1970, Detroit, Mich., and June 8–11, 1970, Brussels, Belgium, Intern. Safety Conf. Compendium No. P–30, 112–132 (Society of Automotive Engineers, Inc., New York, N.Y., June 1970).

Childs, G. E., Diller, D. E., Refractive index of liquid deuterium, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cyrogenic Engineering **15**, Paper No. C–2, 65–69 (Plenum Press, Inc., New York, N.Y., 1970).

Christensen, R. G., Hoeve, C. A. J., Comparison between theoretical and experimental values of the volume changes accompanying rubber extension, J. Polymer Sci. **8**, Part A–1, 1503–1512 (1970).

Clark, A. F., Childs, G. E., Wallace, G. H., Low-temperature electrical resistivity of some engineering alloys, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. C–5, 85–90 (Plenum Press, Inc., New York, N.Y., 1970).

Colwell, J. H., Mangum, B. W., Thornton, D. D., Wright, J. C., Moos, H. W., Low-temperature magnetic properties of $DyPO_4$: An ideal three-dimensional ising antiferromagnet, Phys. Rev. Letters **23**, No. 21, 1245–1247 (Nov. 24, 1969).

Cooper, M. J., Critical indices of the Z–L polymer folding transition, J. Chem. Phys. Letter **52**, No. 1, 473–474 (Jan. 1, 1970).

Coxon, B., Model parameters for the analysis of skew conformations of carbohydrates by p.m.r. spectroscopy, Carbohydrate Res. **13**, 321–330 (1970).

Coxon, B., The synthesis, equilibration, and conformation of diastereoisomeric 1,2-*O*-isopopylidene-3,5-*O*-(methoxymethylidene)-6-*O*-*p*-tolysulfonyl-α-D-glucofuranoses. Conformational evidence from a nuclear Overhauser effect, Carbohydrate Res. **12**, 313–334 (May 15, 1970).

Coyle, T. D., Silica, Kirk-Othmer Encyclopedia of Chemical Technology, 2nd Edition, **18**, 46–61 (1969).

Cuthill, J. R., A soft x-ray spectrometer with improved drive, Rev. Sci. Instr. **41**, No. 3, 422–423 (Mar. 1970).

Daney, D. E., Rapial, A. S., Preparation and characterization of slush hydrogen and nitrogen gels, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. L–4, 467–475 (Plenum Press, Inc., New York, N.Y., 1970).

Diamond, J. J., Editor, Bibliography on the high temperature chemistry and physics of materials, Jan., Feb., Mar., 1970, NBS Spec. Publ. 315–5, 82 pages (Apr. 1970) 75 cents, SD Catalog No. C13.10:315.5.

Diller, D. E., Roder, H. M., Thermal conductivity measurements on fluid hydrogen at 17 to 200 °K and pressures to 10 MN/m^2, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. C–1, 58–64 (Plenum Press, Inc., New York, N.Y., 1970).

DiMarzio, E. A., Guttman, C. M., Separation by flow, Macromolecules **3**, No. 2, 131–146 (Mar.–Apr. 1970).

Duncan, A. G., Hiza, M. J., A multipurpose phase equilibrium apparatus to study mixtures of cryogenic fluids: application to argon-methane (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969 Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. B–2, 42–45 (Plenum Press, Inc., New York, N.Y., 1970).

Eisele, J. W., Technical Standards Coordinator, Horticultural grade perlite, Nat. Bur. Stand. (U.S.), Prod. Stand. 23–70, 9 pages (May 1970) 10 cents, SD Catalog No. C13.20/2:23–70.

Eisele, J. W., Technical Standards Coordinator, Types and sizes of forms for one-way concrete joist construction, Nat. Bur. Stand. (U.S.), Prod. Stand. 16–69, 16 pages (June 1970), 10 cents, SD Catalog No. C13.20/2:16–69.

Eisenhart, C., Anniversaries in 1970 of interest to statisticians, Am. Stat. **24**, No. 1, 25–68 (Feb. 1970).

Engen, G. F , An evaluation of the "back-to-back" method of measuring adapter efficiency, IEEE Trans. Instr. Meas. **IM–19**, No. 1, 18–22 (Feb. 1970).

Evenson, K. M., Wells, J. S., Matarrese, L. M., Absolute frequency measurements of the CO_2 cw laser at 28 THz (10.6μm), Appl. Phys. Letters **16**, No. 6, 251–253 (Mar. 15, 1970).

Evenson, K. M., Wells, J. S., Matarrese, L. M., Elwell, L. B., Absolute frequency measurements of the 28- and 78-μm cw water vapor laser lines, Appl. Phys. Letters **16**, No. 4, 159–161 (Feb. 15, 1970).

Faison, T. K., Jr., Davis, J. C., Achenbach, P. R., Performance of louvered devices as air mixers, NBS Bldg. Sci. Ser. 27, 22 pages (Mar. 1970) 30 cents, SD Catalog No. C13.29/2:27.

Fishman, L., Mountain, R. D., Activity coefficients of solutions from the intensity ratio of Rayleight to Brillouin scattering, J. Phys. Chem. **74**, No. 10, 2178–2182 (May 14, 1970).

Forum, Ohio State Univ., Columbus, Ohio, November 4–6, 1969, 31, 37–45, Porcelain Enamel Institute, Washington, D.C. (1969).

Fraker, A. C., Ruff, A. W., Jr., Microstructural studies of 7075 Al-relation to the directional sensitivity for stress corrosion cracking, Corrosion Sci. **10**, 191–195 (1970).

Franklin, A. D., Bennett, H. S., Editors, ARPA–NBS program of research on high temperature materials and laser materials. Reporting Period July 1 to December 31, 1969, Nat. Bur. Stand. (U.S.), Tech. Note, 531, 75 pages (June 1970) 70 cents, SD Catalog No. C13.46:531.

Frederikse, H. P. R., The electronic band structure of strontium titanate: Theory and experiment (Proc. Conf. International Advanced Study Institute: Electronic Structures in Solids, Chania, Crete, Greece, June 30–July 14, 1968), Chapter in Electronic Structures in Solids, 259–269 (Plenum Press, New York, N.Y., 1969).

Funkhouser, A. T., Mielenz, K. D., High-speed holographic interferrometry, Appl. Opt. **9**, No. 5, 1215 (May 1970).

Furukawa, G. T., Saba, W. G., Sweger, D. M., Plumb, H. H., Normal boiling point and triple point temperatures of neon, Metrologia **6**, No. 1, 35–37 (Jan. 1970).

Gadzuk, J. W., Coupled surface-plasmom modes in metal-thin-film-vacuum sandwiches, Phys. Rev. B. **1**, No. 3, 1267–1269 (Feb. 1, 1970).

Gadzuk, J. W., Resonance transmission in electron emission from surfaces with adsorbed atoms, Surface Sci. **18**, 193–203 (1969).

Gadzuk, J. W., Resonance-tunneling spectroscopy of atoms adsorbed on metal surfaces: theory, Phys. Rev. B. **1**, No. 5, 2110–2129 (March 1, 1970).

Gallagher, A., State-transfer collisions (Proc. Sommerfeld Centennial Memorial Meeting, Munich, Germany, Sept. 9–14, 1968), Chapter in Physics of the One- and Two-Electron Atomics, F. Bopp and H. Kleinpoppen, Eds., pp. 788–794 (North Holland Publ. Co., Amsterdam, The Netherland, 1969).

Geil, G. W., Feinberg, I. J., Temperature changes in specimens in microplasticity tests, Met. Trans. **1**, No. 7, 1845–1851 (July 1970).

Giacchetti, A., Stanley, R. W., Zalubas, R., Proposed secondary-standard wavelengths in the spectrum of thorium, J. Opt. Soc. Am. **69**, No. 4, 474–489 (April 1970).

Ginnings, D. C., Editor, Precision measurement and Calibration. Selected NBS Papers on heat, Nat. Bur. Stand. (U.S.), Spec. Publ. 300, Vol. **6**, 387 pages (February 1970) $5.00, SD Catalog No. C13.10:300/v.6.

Graminski, E. L., The stress-strain behavior of accelerated and naturally aged papers, TAPPI **53**, No. 3, 406–410 (Mar. 1970).

Greenfeld, S. H., Warner, E. R., Reinhart, H. W., Bibliographies on fabric flammability. Part 4. Interior furnishings, Nat. Bur. Stand. (U.S.), Tech. Note 498–1, 24 pages (June 1970) 30 cents, SD Catalog No. C13.46:498–1.

Greenspan, M., Tschiegg, C. E., Cavitation nucleated by $^{10}B(n, \alpha)^7Li$, Nucl. Instr. Methods **82**, No. 3, 310–312 (1970).

Goltz, S. M., Tanner, C. B., Thurtell, G. W., Jones, F. E., Evaporation measurements by an Eddy correlation method, Water Resources Res. **6**, No. 2, 440–446 (Apr. 1970).

Gonano, R., Adams, E. D., In situ vapor pressure measurement for low temperature thermometry, Rev. Sci. Instr. **41**, No. 5, 716–719 (May 1970).

Gross, D., Natrella, M. G., Interlaboratory comparison of the potential heat test method, Am. Soc. Testing Mater. Spec. Tech. Publ. 464, Fire Test Performance, 127–152 (1970).

Gugeler, A. L., A study of the adherence of porcelain enamel to aluminum, Chapter in Proc. of the Procelain Enamel Institute.

Guildner, L. A., Stimson, H. F., Edsinger, R. E., Anderson, R. L., An accurate mercury manometer for the NBS gas thermometer, Metrologia **6**, No. 1, 1–18 (Jan. 1970).

Haar, L., Sengers, J. M. H. L., The solubility of condensed substances in dense gases, and the effect on PVT properties, J. Chem. Phys. **52**, No. 10, 5069 (May 15, 1970).

Haber, S., Sequences of numbers that are approximately completely equidistributed, J. Assoc. Computing Mach. **17**, No. 2, 269–272 (Apr. 1970).

Hamer, W. J., DeWane, H. J., Electrolytic conductance and the conductances of the halogen acids in water, Nat. Stand. Ref. Data Ser., Nat. Bur. Stand. (U.S.), 33, 37 pages (May 1970) 50 cents, SD Catalog No. C13.48:33.

Hamilton, W. C., Edmonds, J. W., Tippe, A., Rush, J. J., Methyl group rotation and the low temperature transition in hexamethylbenzene. A neutron diffraction study, Discussions Faraday Soc., No. 48, 192–204 (1969).

Handy, L. B., Brinckman, F. E., Chemistry of the methoxyfluorotunsten(VI) series, Chem. Commun. 214–215 (1970).

Hanson, D. W., Quasi-optical components using total reflection in dielectrics, IEEE Trans. Microwave Theory Tech. **MTT–18**, No. 4, 233–234 (April 1970).

Harvey, J. L., Brenner, F. C., Tire use survey. The physical condition, use, and performance of passenger car tires in the United States of America, NBS Tech. Note 528, 59 pages (May 1970), 60 cents, SD Catalog No. C13.46:528.

Heinrich, K. F. J., Electron and ion microprobe analysis (Proc. 10th Symp. Electron, Ion, and Laser Beam Technology, Gaithersburg, Md., May 21–23, 1969), Chapter in Record of the 10th Symposium on Electron, Ion and Laser Beam Technology, L. Marton, Ed., 353–360 (San Francisco Press, San Francisco, Calif., 1969).

Heinrich, K. F. J., Present state of the classical theory of quantitative electron probe microanalysis, Nat. Bur. Stand. (U.S.), Tech. Note 521, 17 pages (Aug. 1970) 30 cents, SD Catalog No. C13.46:521.

Henrich, K. F. J., Yakowitz, H., Quantitative electron probe microanalysis: Uncertainty in the atomic number correction, Mikrochim. Acta **7**, No. 1, 123–134 (Jan. 1970).

Hellwig, H., Hydrogen spin exchange frequency shifts, NBS Tech. Note 387, 13 pages (March 1970), 30 cents, SD Catalog No. C13.46:387.

Hendricks, R. C., Simoneau, R. J., Smith, R. V., Survey of heat transfer to near-critical fluids (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. G–1, 197–237 (Plenum Press, Inc., New York, N.Y., 1970).

Heydemann, P. L., Ultrasonic measurements at pressures up to 50 kbar, Proc. Intern. Collog. Solids at High Pressures, Grenoble, France, Sept. 8, 1969, No. 188, pp. 461–467 (Centre National de la Recherche Scientifique, Paris, France, 1970).

Hiza, M. J., Solid-vapour equilibria research on systems of interest in cryogenics, Cryogenics **10**, No. 2, 106–115 (April 1970).

Horn, W. A., Optimal networks joining n points in the plane (Proc. IVth International Symp. The Theory of Traffic Flow, Karlsruhe, Germany, June 18–20, 1968), Chapter in Strassenbau und Strassenverkehrstechnik 86, W. Leutzbach and P. Baron, Ed., 160–166 (Bundesminister fur Verkehr, Germany, 1969).

Hougen, J. T., The calculation of rotational energy levels and rotational line intensities in diatomic molecules, Nat. Bur. Stand. (U.S.), Monogr. 115, 52 pages (June 1970) 55 cents, SD Catalog No. C13.44:115.

Hummer, D. G., Spectral line formation by noncoherent scattering with a dipole phase-function, Astrophysical Letters **5**, 1 (1970).

Hummer, D. G., Observatory Report—Joint Institute for Laboratory Astrophysics of the National Bureau of Standards and the University of Colorado, Bull. Am. Astron. Soc. **2**, No. 1, 59–65 (1970).

Jones, M. C., Far infrared absorption in liquefied gases, NBS Tech. Note 390, 39 pages (April 1970), 50 cents, SD Catalog No. C13.46:390.

Jones, R. N. Huntley, L. E., A precision, high frequency calibration facility for coaxial capacitance standards, NBS Tech. Note 386, 27 pages (March 1970), 35 cents, SD Catalog No. C13.46:386.

Kasen, M. B., Grain boundary resistivity of aluminum, Phil. Mag. **21**, No. 171, 599–610 (Mar. 1970).

Kidnay, A. J., Hiza, M. J., Dickson, P. F., Adsorption kinetics in a ternary system containing hydrogen (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering **15**, Paper No. B–3, 46–49 (Plenum Press, Inc., New York, N.Y., 1970).

Kokoszka, G. F., Brinckman, F. E., Electron paramagnetic resonance studies of phosphorus-containing reactive intermediates, J. Am. Chem. Soc. **92**, No. 5, 1199–1205 (1970).

Kose, V. E., Sullivan, D. B., Influence of external noise on microwave-induced Josephson steps, J. Appl. Phys. **41**, No. 1, 169–174 (Jan. 1970).

Kuriyama, M., The dynamical scattering amplitude of an imperfect crystal, Acta Cryst. **A26**, No. 6, 56–59 (Jan. 1970).

Kushner, L. M., Are there standards? (Proc. National Safety Congress Conf., Chicago, Ill., Oct. 28, 1969), National Safety Congress Trans., Civic Leadership **6**, 12–16 (1969).

Lafferty, W. J., Maki, A. G., Coyle, T. D., High resolution infrared spectrum and structure of diborane, J. Mol. Spectr. **33**, No. 2, 345–367 (1970).

LaFleur, P. D., Editor, Activation analysis section: Summary of activities July 1968 to June 1969, Nat. Bur. Stand. (U.S.) Tech. Note 508, 147 pages (July 1970) $1.25, SD Catalog No. C13.46:508.

Landgrebe, W. B., Mann, W. B., Schima, F. J., The distribution of 70-kV prasodymium ions in iron foil and a technique for removing thin layers of iron, Intern. J. Appl. Radiation Isotopes **21**, No. 3, 169–171 (Mar. 1970).

Larson, W., Desch, R. F., Gillard, B. F., Further analysis of the off-null versus power ratio method of attenuation measurement, IEEE Trans. Microwave Theory Tech. **MTT–18**, No. 2, 112–113 (Feb. 1970).

Lawton, R. A., Allred, C. M., Hudson, P. A., A wide-range CW power measurement technique, IEEE Trans. Instr. Meas. **IM–19**, No. 1, 28–34 (Feb. 1970).

Lawton, R. A., New standard of electric field strength, IEEE Trans. Instr. Meas. **IM–19**, No. 1, 45–51 (Feb. 1970).

Levin, E. M., X-ray determination of the thermal expansion of silver nitrate, J. Am. Ceram. Soc. **52**, No. 1, 53–54 (January 1969).

Locke, J. W., Penn, R., Rick, J., Bunten, E., Hare, G., Compilation and use of criminal court data in relation to Pre-Trial Release of defendants: pilot study, Nat. Bur. Stand. (U.S.), Tech. Note 535, 242 pages (Aug. 1970) $1.75, SD Catalog No. C13.46:535.

Lutz, G. J., Boreni, R. J., Maddock, R. S., Meinke, W. W., Editors, Activation Analysis: A bibliography, NBS Tech. Note 467, Part 1, Addendum 1, 176 pages (Dec. 1969) and Part 2, Appendices, Revision 1, NBS Tech. Note 467, 264 pages (Dec. 1969), $8.50 per set of 3 volumes, sold in sets only. Part 1 of September 1968 accompanies these sets, SD Catalog No. C13.46:467.

Lutz, G. J., Determination of oxygen in sodium by photon activation analysis, Anal. Chem. **42**, No. 4, 531–532 (Apr. 1970).

Lutz, G. J., Editor, Determination of the light elements in metals: A bibliography of activation analysis papers, NBS Tech. Note 524, 66 pages (May 1970), 75 cents, SD Catalog No. C13.46:524.

Lutz, G. J., Editor, 14-MeV neutron generators in activation analysis: a bibliography, Nat. Bur. Stand. (U.S.), Tech. Note 533, 91 pages (June 1970) $1.00, SD Catalog No. C13.46:533.

Lutz, G. J., Editor, Oceanography: a bibliography of selected activation analysis literature, Nat. Bur. Stand. (U.S.), Tech. Note 534, 36 pages (June 1970) 50 cents, SD Catalog No. C13.46:534.

Lutz, G. J., Editor, Pollution analysis: a bibliography of the literature of activation analysis, Nat. Bur. Stand. (U.S.), Tech. Note 532, 32

pages (June 1970) 45 cents, SD Catalog No. C13.46:532.

Lutz, G. J., Masters, L. W., Determination of carbon in high purity metals by photon activation analysis, Anal. Chem. **42,** No. 8, 948–950 (July 1970).

Mackay, D. R., American Softwood Lumber Standard, NBS Prod. Stand. 20–70, 26 pages (Jan. 1970), 35 cents, SD Catalog No. C13.20/2:20–70.

Machta, L., Hughes, E., Atmospheric oxygen in 1967–1970, Science **168,** No. 3939, 1582–1584 (June 26, 1970).

Madey, T. E., Yates, J. T., Jr., Chemisorption on single crystals: H_2 on (100) tungsten, Proc. Colloque Intern. sur la Structure et les Proprietes des Surfaces des Solides, Paris, France, July 7–11, 1969, No. 187, pp. 155–162 (Centre National de la Recherche Scientifique, Paris, France, 1970).

Madey, T. E., Yates, J. T., Jr., King, D. A., Uhlaner, C. J., Isotope effect in electron stimulated desorption: oxygen chemisorbed on tungsten, J. Chem. Phys. **52,** No. 10, 5215–5220 (May 15, 1970).

Mahler, R. J., Nominal $\Delta m = \pm 1, 2, 3, 4, 5$ magnetic nuclear transition probabilities in ruby, Phys. Rev. **188,** No. 2, 584–590 (Dec. 1969).

Mahler, R. J., James, L. W., Acoustic and magnetic effects involving the 19F nuclei in antiferromagnetic KMnF, J. Appl. Phys. **41,** No. 4, 1633–1636 (March 15, 1970).

McKinney, J. E., Oser, H. J., Stability of acoustic waves within a viscous compressible heat-conduction fluid, J. Acoust. Soc. Am. **47,** No. 3, 781–785 (March 1970).

Mclaughlin, W. L., Photographic film dosimeters, Chapter in Manual on Radiation Dosimetry, N. W. Holm and R. J. Berry, Eds., 387–394 (Marcel Dekker, Inc., New York, N.Y., 1970).

McLaughlin, W. L., Radiochromic dye-cyanide dosimeters, Chapter in Manual on Radiation Dosimetry, N. W. Holm and R. J. Berry, Eds., 377–385 (Marcel Dekker, Inc., New York, N.Y., 1970).

McNish, A. G., Progress on the Metric Study Program, Proc. Conf. 18th Annual Meeting of Standards Engineers Society, Washington, D.C., Sept. 17–19, 1969, pp. 53–56 (Standards Engineers Society, Philadelphia, Pa., 1969).

McNish, A. G., The impact of increasing metric usage, Proc. Conf. Industrial Heating Equipment Association, Spring (77th Consecutive) Meeting, Hot Springs, Virginia, May 19–21, 1969, Exhibit E (Industrial Heating Equipment Assoc., Washington, D.C., 1969).

Margoshes, M., Data acquisition and computation in spectrochemical analysis: a forecast, Spectrochim. Acta **25B,** 113–122 (June 15, 1969).

Margoshes, M., Scribner, B. F., Emission spectrometry, Anal. Chem. **42,** No. 5, 398R–417R (April 1970).

Marton, L., Survey of microanalysis-interpolation and extrapolation, Chapter in Electron Probe Microanalysis, pp. 1–14 (Academic Press, Inc., New York, N.Y., 1969).

Marvin, R. S., Rheological Models and measurements, (Proc. 5th. Intern. Congress on Rheology, Kyoto, Japan, Oct. 7–11, 1968), Chapter in Fifth International Congress on Rheology, S. Onogi, Ed., 1, 85–96, (University of Tokyo Press and University Park Press, Tokyo, Japan, 1969).

Mauvais, C. J., Latanision, R. M., Ruff, A. W., Jr., On the anisotropy observed during the passivation of nickel monocrystals, J. Electochem. Soc. **117,** No. 7, 902–903 (July 1970).

Mazur, J., Higher order Markov chains and statistical thermodynamics of linear polymers, Chapter 6 in Markov Chains and Monte-Carlo calculations in Polymer Science, G. G. Lowry, Ed., 153–185 (Marcel Dekker, Inc., New York, N.Y., 1970).

Melmed, A. J., Carroll, J. J., Ellipsometry, LEED, and FEM study of evaporated epitaxial films of iron on (011) tungsten, Surface Sci. Letters to the Editor **19,** No. 1, 234–248 (January 1970).

Messina, C. G., Hilsenrath, J., Edit-insertion programs for automatic typesetting of computer printout, NBS Tech. Note 500, 50 pages (Apr. 1970), 60 cents, SD Catalog No. C13.46:500.

Michaels, S., Green, M. S., Larsen, S. Y., Equilibrium critical phenomena in fluids and mixtures: a comprehensive bibliography with key-word descriptors, Nat. Bur. Stand. (U.S.), Spec. Publ. 327, 235 pages (June 1970) $4.00, SD Catalog No. C13.10:327.

Mighell, A. D., Reimann, C. W., Santoro, A., The crystal structure of the 2:1 dimer $[(\text{pyridine N-oxide})_2CuBr_2]_2$, Chem. Commun. **4,** 204 (1970).

Milligan, D. E., Jacox, M. E., Infrared and ultraviolet spectra of the products of the vacuum-ultraviolet photolysis of silane isolated in an argon matrix, J. Chem. Phys. **52,** No. 5, 2594–2608 (Mar. 1, 1970).

Milligan, D. E., Jacox, M. E., Guillory, W. A., Infrared spectrum of NO_2^-ion isolated in an argon matrix, J. Chem. Phys. **52,** No. 8, 3864–3868 (April 15, 1970).

Molino, J. A., Noise the growing hazard, Dialog **1,** No. 4, 26–28 (1970).

Mountain, R. D., Comments on fluctuation theory calculations of the specific heat near the critical point, J. Chem. Phys. **52,** No. 7, 3852–3853 (April 1, 1970).

Mountain, R. D., Liquids: Dynamics of liquid structure, Critical Reviews in Solid State Sciences **1,** No. 1, 5–46 (Chemical Rubber Co., Cleveland, Ohio, March 1970).

Nargolwalla, S. S., Crambes, M. R., Suddueth, J. E., Photon self-adsorption corrections for the minimization of systematic errors in 14-MeV neutron activation analysis, Anal. Chim. Acta **49,** No. 3, 425–436 (Mar. 1970).

Nargolwalla, S. S., LaFleur, P. D., Characterization of standard reference materials for industry and research by neutron activation analysis, Proc. American Nuclear Society Topical Meeting, San Juan, Puerto Rico, May 4–6, 1969, 183–186 (1969).

Negas, T., Roth, R. S., The system $SrMnO_{3-x}$, J. Solid State Chem. **1,** 409–418 (1970).

Newell, K. G., Jr., Technical Standards Coordinator, Acrylonitrile-butadiene-styrene (ABS) plastic pipe (standard dimension ratio), Nat. Bur. Stand. (U.S.), Prod. Stand. 19–69, 13 pages (June 1970) 15 cents, SD Catalog No. C13.20/2:19–69. Supersedes NBS Commercial Standard CS254–63.

Newell, K. G., Jr., Technical Standards Coordinator, Acrylonitrile-butadiene-styrene (ABS) plastic pipe (schedules 40 and 80), Nat. Bur. Stand. (U.S.), Prod. Stand. 18–69, 15 pages (June 1970) 15 cents, SD Catalog No. C13.20/2:18–69. Supersedes NBS Commercial Standard CS218–59.

Newell, K. G., Technical Standards Coordinator, Poly(vinyl chloride) (PVC) plastic pipe (schedules 40, 80, and 120), Nat. Bur. Stand. (U.S.), Prod. Stand. 21–70, 16 pages (June 1970) 15 cents, SD Catalog No. C13.20/2:21–70.

Newell, K. G., Technical Standards Coordinator, Poly(vinyl chloride) (PVC) plastic pipe (standard dimension ratio), Nat. Bur. Stand. (U.S.), Prod. Stand. 22–70, 14 pages (June 1970) 20 cents, SD Catalog No. C13.20/2:22–70.

Newman, M., A table of the first factor for prime cyclotomic fields, Math. Compt. **24,** No. 109, 215–219 (Jan. 1970).

Page, C. H., Relations among systems of electromagnetic equations, Am. J. Phys. **38,** No. 4, 421–424 (April 1970).

Page, C. H., The trolley problem. Am. J. Phys. **38,** No. 5, 608–610 (May 1970).

Parke, W. C., Ponzini, R. G., Williams, H. T., Remarks on finite-energy sum rules, Phys. Rev. **1,** No. 1–D, 134–139 (Jan. 1, 1970).

Paulsen, P. J., Alvarez, R., Mueller, C. W., Spark source mass spectrographic analysis of ingot iron for silver, copper, molybdenum and nickel by isotope dilution and for cobalt by an internal standard technique, Anal. Chem. **42,** 673–675 (May 1970).

Penner, S., Beam handling techniques for electron linear accelerators, NBS Tech. Note 522, 21 pages (Apr. 1970), 30 cents, SD Catalog No. C13.46:522.

Pfeiffer, E. R., Schooley, J. F., Effect of stress on the superconducting transition temperature of $SrTiO_3$, J. Low Temp. Phys. **2,** No. 3/4, 333–352 (August 7, 1970).

Phillips, S. L., Johnson, C. E., Effect of atmospheric exposure on the contact resistance of selected tin alloys, J. Electrochem. Soc. **117,** No. 6, 827–830 (June 1970).

Photonuclear data index, Jan. 1965 through Jan. 1970, NBS Spec. Publ. 322, 86 pages (May 1970), 75 cents, SD Catalog No. C13.10:322, (Supersedes Supplement 1 to Misc. Publ. 277).

Phucas, C. B., Technical Standards Coordinator, Salt packages, Nat. Bur. Stand. (U.S.), Prod. Stand. 14–69, 9 pages (May 1970) 10 cents, SD Catalog No. C13.20/2:14–69. (Supersedes Simplified Practice Recommendation 70–54).

Plummer, E. W., Young, R. D., Field-emission studies of electronic energy levels of adsorbed atoms, Phys. Rev. B. **1,** No. 5, 2088–2109 (March 1, 1970).

Powell, F. X., Johnson, D. R., Microwave detection of $H_2^{18}O$, Phys. Rev. Letters **24,** No. 12, 637 (Mar. 23, 1970).

Powell, R. L., Cryogenic materials needs, Proc. 5th Annual Natl. Conf. Industrial Research, Sept. 18–19, 1969, Chicago, Ill., Chapter in Applying Emerging Technologies, pp. 102–107 (Industrial Research, Inc., Beverly Shores, Indiana, 1970).

406-680 O - 71 - 7

Qureshi, I. H., McClendon, L. T., LaFleur, P. D., Extraction studies of the group IIIB–VIIB elements and the lanthanides utilizing bis(2-ethyl-hexyl) orthophosphoric acid, Radiochim. Acta. **12,** No. 2, 107–111 (Aug. 1969).

Randall, E. B., Jr., Lashof, T. W., An interlaboratory study of the measurement of the bursting strength of paper, Tappi **53,** No. 5, 799–809 (May 1970).

Raveche, H. J., Mayer, J. E., Theory of diffusion in simple liquid mixtures, J. Chem. Phys. **52,** No. 8, 3990–4004 (April 15, 1970).

Reed, R. P., Durchol, R. L., Cryostat and strain measurement for tensile tests to 1.5 °K (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. D–3, 109–116 (Plenum Press, Inc., New York, N. Y., 1970).

Ritter, J. J., Coyle, T. D., Halogen exchange reactions of diboron tetrahalides, J. Chem. Soc. Sec. **A,** pp. 1303–1304 (1970).

Robbins, D., Englert, A. J., Editors, Disclosures on: Viscous damped wind vane; nonskid road or runway; regrooved pneumatic tire with removable inserts; device for radially positioning a rotating wheel; method for fabricating precision waveguide sections; distortion-cancelling loudspeaker system; cryogenic fluid density measurement system; and controlled-atmosphere weathering device, Nat. Bur. Stand. (U.S.), Tech. Note 536, 25 pages (June 1970) 35 cents, SD Catalog No. C13.46:536.

Robertson, A. F., Gross, D., Fire load, fire severity, and fire endurance, Am. Soc. Testing Mater. ASTM Spec. Tech. Publ. **464,** Fire Test Performance, pp. 3–29 (January 1970).

Rockett, J. A., Objectives and pitfalls in the simulation of building fires with a computer, Fire Tech. **5,** No. 4, 311–322 (November 1969).

Rogers, E. E., Abramowitz, S., Jacox, M. E., Milligan, D. E. Matrix-isolation studies of the infrared spectra of the free radicals CCl_3 and CBr_3, J. Chem. Phys. **52,** No. 5, 2198–2204 (Mar. 1, 1970).

Romanoff, M., Corrosion evaluation of steel test piles exposed to permafrost soils, Proc. National Association of Corrosion Engineers 24th Conf., Houston, Texas, March 10–14, 1969, 6–13 (National Association of Corrosion Engineers, Houston, Texas, 1970).

Rosenblatt, J. R., Joiner, B. L., Hogben, D., OMNITAB-Rapid statistical manipulation (Proc. Conf. Small-Area Statistics, American Statistical Association, New York, N. Y., Aug. 21, 1969), Paper in Census Tract Papers Final 1970 Census Plans and Four Programing Systems for Computerized Data Retrieval and Manipulation, Series GE–40, No. 6, pp. 46–59 (Superintendent of Documents, Government Printing Office, Washington, D.C., 20402, 65 cents, 1970).

Ruegg, F. C., Spijkerman, J. J. DeVoe, J. R., A Mössbauer spectrometer for the structural analysis of materials (Proc. Symp. Radioisotope Instruments in Industry and Geophysics, Warsaw, Poland, Oct. 18–22, 1965), Chapter in Radioisotope Instruments in Industry and Geophysics I, 325–335 (International Atomic Energy Agency, Vienna, Austria, May 1966).

Ryan, J. V., Implementation of the Flammable Fabrics Act, Proc. Conf. Flame Retardant Cotton Batting Workshop, New Orleans, Louisiana, Nov. 15, 1968, ARS 72–72, 1–77 (Oct. 1969).

Santoro, A., Mighell, A. D., Determination of reduced cells, Acta. Cryst. **A26,** No. 6, 124–127 (Jan. 1970).

Schneider, W. E., A one-solar-constant irradiance standard, Appl. Opt. **9,** No. 6, 1410–1418 (June 1970).

Schroeder, L. W., Infrared and raman spectra of $ClHCl^-$ and $BrHBr^-$ ions in $CsCl \cdot k/3H_3OHCl_2$ and $C_sBr \cdot 1/3H_3OHBr_2$, J. Chem. Phys. **52,** No. 4, 1972–1978 (Feb. 1970).

Schwerdtfeger, W. J., Corrosion rates of metals, alloys, and galvanic couples as measured by polarization techniques, Mater. Res. Std. **10,** No. 3, 22–28 (Mar. 1970).

Searles, S. K., Sieck, L. W., Ausloos, P., Reactions of $C_2H_6^+$. Formation of the $(C_2H_6)_2^+$ ion, J. Chem. Phys. **53,** No. 2, 849–850 (July 15, 1970).

Searles, S. K., Sieck, L. W., High pressure photoionization mass spectrometry, III. Reactions of $NO^+(X^1\Sigma^+)$ with C_3-C_7 hydrocarbons at thermal kinetic energies, J. Chem. Phys. **53,** No. 2, 794–797 (July 15, 1970).

Sher, A. H., Coleman, J. A., Lithium driftability in detector-grade germanium, IEEE Trans. Nucl. Sci. **NS–17,** No. 3, 125–129 (June 1970).

Sher, A. H., Nomographs for use in the fabrication and testing of Ge(Li) detectors, Nat. Bur. Stand. (U.S.), Tech. Note 537, 18 pages (Aug. 1970) 30 cents, SD Catalog No. C13.46:537.

Sher, A. H., Keery, W. J., Variation in the effective Fano factor in a Ge(Li) detector, IEEE Trans. Nucl. Sci. **NS–17,** No. 1, 39–43 (February 1970).

Shields, W. R., Editor, Analytical Mass Spectrometry Section: Summary of activities July 1968 to June 1969, NBS Tech. Note 506, 55 pages (May 1970), 55 cents, SD Catalog No. C13.46:506.

Shinyayev, A. Y., Butrymowicz, D. B., Interdiffusion in and the phase diagram for vanadium-rich alloys of the V–Al system at pressures 0 to 47 kbar, Met. Trans. **1,** No. 7, 1905–1907 (July 1970).

Shipp, W. E., Surface tension of binary mixtures of several organic liquids at 25 °C, J. Chem, Eng. Data **15,** No. 2, 308–311 (April 1970).

Simson, B. G., Mandel, J., Brenner, F. C., Research for a uniform quality grading system for tires. III. Breaking energy Rubber Chem. and Tech. **43,** No. 2, 356–369 (March 1970).

Sindt, C. F., Ludtke, P. R., Slush hydrogen flow characteristics and solid fraction upgrading (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. I–6, 382–390 (Plenum Press, Inc., New York, N.Y., 1970).

Smith, S. J., Survey on electron-atom collision experiments (Proc. Sommerfeld Centennial Memorial Meeting, Munich, Germany, Sept. 9–14, 1968), Chapter in Physics of the One- and Two-Electron Atoms, F. Bopp and H. Kleinpoppen, Eds., pp. 574–597 (North Holland Publ. Co., Amsterdam, The Netherlands, 1969).

Snediker, D. K., The Mössbauer effect of Sn^{119} in palladium-rich palladium-tin solid solutions (Proc. 2nd Symp. Mössbauer Effect Methodology, New York, N.Y., Jan. 25, 1966), Chapter in Mössbauer Effect Methodology **2,** 161–170 (Jan. 1966).

Snelleman, W., The measurement and calculation of flame temperatures, Chapter 7 in Flame Emission and Atomic Absorption Spectrometry I, 213–239 (Marcel Dekker, Inc., New York, N.Y., 1969).

Snyder, N. S., Heat transport through helium II: Kapitza conductance, Cryogenics **10,** No. 2, 89–95 (April 1970).

Stevens, M. E., Automatic Indexing: A state-of-the-art report, NBS Mono. 91, 290 pages (Mar. 1970), $2.25, SD Catalog No. C13.44:91. (Reissued with additions and corrections Feb. 1970).

Stevens, M. E., Research and development in the computer and information sciences. 2. Processing, storage, and output requirements in information processing systems: A selective literature review, NBS Mono. 113 Vol. 2, 125 pages (May 1970), $1.25, SD Catalog No. C13.44:113/Vol 2.

Stevens, M. E., Technical Standards Coordinator, Research and development in the computer and information sciences. 3. Overall system design considerations a selective literature review, Nat. Bur. Stand. (U.S.), Monogr. 113, Vol. 3, 147 pages (June 1970) $1.25, SD Catalog No. C13.44:113, Vol. 3.

Stevenson, D. R., Technical Standards Coordinator, Polyethylene sheeting (construction, industrial, and agricultural applications), Nat. Bur. Stand. (U.S.), Prod. Stand. 17–69, 20 pages (June 1970) 15 cents, SD Catalog No. C13.20/2:17–69 (Supersedes Commerical Standard CS 238–61).

Stevenson, D. R., Technical Standards Coordinator, Custom contact-molded reinforced-polyester chemical-resistant process equipment, Nat. Bur. Stand. (U.S.), Prod. Stand. 15–69, 28 pages (June 1970) 20 cents, SD Catalog No. C13.20/2:15–69.

Steward, W. G., Smith, R. V., Brennan, J. A., Cooldown transients in cryogenic transfer lines (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. I–2, 354–363 (Plenum Press, Inc., New York, N.Y., 1970).

Stief, L. J., Payne, W. A., DeCarlo, V. J., Gorden, R., Jr., Ausloos, P., Comment on "Xenon sensitized photolysis of carbon dioxide", J. Chem. Phys. **53,** No. 1, 475–476 (July 1, 1970).

Straty, G. C., Prydz, R., Melting curve and triple-point properties of fluorine, Physics Letters **31A,** No. 6, 301–302 (March 23, 1970).

Straty, G. D., Prydz, R., The vapor pressure of liquid fluorine, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. B–1, 36–41 (Plenum Press, Inc., New York, N.Y., 1970).

Sugar, J., Configuration $4f^{13}5d$ of doubly ionized ytterbium, J. Opt. Soc. Am. **60,** No. 4, 571–572 (April 1970).

Sugar, J., Spectrum of doubly ionized thulium (Tm III), J. Opt. Soc. Am. **69,** No. 4, 454–466 (April 1970).

Swartzendruber, L. J., Bennett, L. H., Clustering, cold work, and

the Mössbauer effect doublet structure in Cu-Ni-Fe alloys, Physics Letters **31A,** No. 10, 581–582 (May 18, 1970).

Swiggard, E. M., Clabaugh, W. S., Preparation of barium titanate semiconductors containing controlled amounts of neodymium, Am. Ceram. Soc. Bull. **45,** No. 9, 777–781 (Sept. 1966).

Thurber, W. R., Determination of oxygen concentration in silicon and germanium by infrared absorption, Nat. Bur. Stand. (U.S.), Tech. Note 529, 21 pages (May 1970) 30 cents, SD Catalog No. C13.46:529.

Tighe, N. J., Hockey, B. J., Ion thinning of electron microscope specimens (Proc. 10th Symp. Electron, Ion and Laser Beam Tech., May 21–23, 1969, National Bureau of Standards, Gaithersburg, Md.), Record of 10th Symposium on Electron, Ion, and Laser Beam Technology, L. Marton, Ed., pp. 375–380 (San Francisco Press, Inc., San Francisco, Calif., 1970).

Torrance, K. E., Orloff, L., Rockett, J. A., Experiments on natural convection in enclosures with localized heating from below, J. Fluid Mech. **36,** Part I, 21–31 (1969).

Torrance, K. E., Orloff, L., Rockett, J. A., Numerical study of natural convection in an enclosure with localized heating from below—creeping flow to the onset of laminar instability, J. Fluid Mech. **36,** Part I, 33–54 (June 1969).

Venable, W. H., Jr., Effects upon radiant intensity measurements due to scattering by optical elements, Appl. Opt. **9,** No. 3, 609–615 (Mar. 1970).

Verdier, P. H., Relaxation behavior of the freely jointed chain, J. Chem. Phys. **52,** No. 11, 5512–5517 (June 1, 1970).

Vicentini-Missoni, M., Joseph, R. I., Green, M. S., Sengers, J. M. H. L., Scaled equation of state and critical exponents in magnets and fluids, Phys. Rev. B. **1,** No. 5, 2312–2331 (March 1, 1970).

Vidal, C. R., Cooper, J., Smith, E. W., Hydrogen stark broadening calculations with the unified classical path theory, Nat. Bur. Stand. (U.S.), Monogr. 116, 143 pages (May 1970) $1.50, SD Catalog No. C13.44:116.

Viezbicke, P. P., Editor, NBS Frequency and time broadcast services, Radio stations WWV, WWVH, WWVB, and WWVL, Nat. Bur. Stand. (U.S.), Spec. Publ. 236, 16 pages (1970) 25 cents, SD Catalog No. C13.11:236 (Supersedes NBS Special Publication 236, 1969 and previous editions).

Weber, L. A., The P–V–T surface of oxygen in the critical region; densities of saturated liquid and vapor, (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. B–4, 50–57 (Plenum Press, Inc., New York, N. Y., 1970).

Weisman, H. M., In the beginning was the word—information needs and uses in science and technology, Paper E 4–2 in Proceedings of the Seventeenth International Technical Communications Conference, Minneapolis, Minn., June 3–6, 1970, B. F. Hylen and R. R. McDaniel, Eds., pp. 1–5 (Society of Technical Writers and Publishers Inc., 1010 Vermont Ave., N. W., Washington, D. C. 20005, 1970).

Weiss, A. W., Krauss, M., Bound-state calculation of scattering resonance energies, J. Chem. Phys. **52,** No. 9, 4363–4368 (May 1, 1970).

Wells, J. S., Evenson, K. M., A new LEPR spectrometer, Rev. Sci. Instr. **41,** No. 2, 226–227 (Feb. 1970).

West, E. D., Jennings, D. A., Power measurement of large laser beams with a small dual-cone calorimeter, Rev. Sci. Instr. **41,** No. 1, (Jan. 1970).

White, H. J., Jr., Counties and county equivalents of the states of the United States, Nat. Bur. Stand. (U.S.), Fed. Info. Process. Stand. Publ. (FIPS Pub.) 6–1, 34 pages (1970) 45 cents, SD Catalog No. C13.52:6–1. Supersedes FIPS PUB 6, Nov. 1, 1968.

White, H. J., Jr., States and outlying areas of the United States, Nat. Bur. Stand. (U.S.), Fed. Info. Process. Stand. Publ. (FIPS Pub) 5–1, 4 pages (1970) 20 cents, SD Catalog No. C13.52:5–1. Supersedes FIPS PUB 5, Nov. 1, 1968.

Wilson, W. K., Development of specifications for archival record materials, Am. Archivist **33,** No. 2, 219–223 (April 1970).

Wilson, W. K., Record papers and their preservation, Chemistry **43,** No. 3, 8–12 (Mar. 1970).

Wims, A. M., Sengers, J. V., McIntyre, D., Shereshefsky, J., Interfacial tension of 3-methylpentane-nitroethane near the critical point, J. Chem. Phys. **52,** No. 6, 3042–3049 (Mar. 15, 1970).

Winogradoff, N. N., Neill, A. H., Petrescu-Prahova, J. B., Band tailing and the spontaneous spectra of compensated epitaxial GaAs laser junctions, IEEE J. Quantum Electron. Q.E. **6,** No. 6, 305–310 (June 1970).

Winogradoff, N. N., Neill, A. H., Jr., Temperature dependence of the power output of the spontaneous emission from GaAs laser diodes, Intern. J. Electron. **28,** No. 5, 401–406 (1970).

Wolfe, W. C., Performance tests for floor coverings, Mater. Res. Stand. **10,** No. 7, 15–18 (July 1970).

Wood, L. A., Creep of "pure-gum" vulcanisates of natural rubber, (Proc. Intern. Natural Rubber Conf., Kuala Lumpur, Malaya, Sept. 1968), J. Rubber Inst. Malaya **22,** No. 3, 309–316 (1969).

Wood, L. A., Shouse, P. J., Passaglia, E., How just appreciable fading of AATCC L–4 wool standard is related to exposure time, Textile Chem. Colorist **2,** No. 11, 182–190 (June 3, 1970).

Yakowitz, H., Comments on "application of color photography to electron microbeam probe samples", Rev. Sci. Instr. **41,** No. 4, 601 (April 1970).

Younglove, B. A., Measurements of the dielectric constant of saturated liquid oxygen (Proc. 1969 Cryogenic Engineering Conf., June 16–18, 1969, Univ. of California, Los Angeles, Calif.), Chapter in Advances in Cryogenic Engineering 15, Paper No. C–3, 70–75 (Plenum Press, Inc., New York, N.Y., 1970).

Zielinzki, W. L., Freeman, D. H., Martire, D. E., Chow, L. C., Gas chromatography and thermodynamics of divinylbenzene separations on 4,4'-dihexoxyazoxybenzene liquid crystal, Anal. Chem. **42,** No. 2, 176–180 (Feb. 1970).

**Publications for which a price is indicated are available by purchase from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (foreign postage, one-fourth additional). The NBS non-periodical series are also available from the National Technical Information Service (NTIS), Springfield, Va. 22151. Reprints from outside journals and the NBS Journal of Research may often be obtained directly from the authors.*

JOUPNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 74B, No. 4, October–December 1970

INDEX TO VOLUME 74B

January to December 1970

A

Page

Absorption 67
Acoustic propagation and stability within an inviscid heat-conducting fluid 67
Analysis of variance 149
Approximations 211
Asymptotic expansions 187
Atchison, T. A., Selecting nonlinear transformation for the evaluation of improper integrals 183
Automatic computing methods for special functions 211

B

Bernstein polynomials 47
Bessel functions 85, 187
Bounded index 55
Bounds 47
Bounds on a polynomial 47
Brualdi, R. A., Newman, M., An enumeration problem for a congruence equation 37

C

Capobianco, M. F., Some theorems on tensor composite graphs 235
Cezairliyan, A., Derivatives of the Grüneisen and Einstein functions 175
Circulants 37
Coefficient algorithms 155
Combinatorics 245
Computer programs 211
Computing pitfalls 211
Confluent hypergeometric functions 85
Congruences 1, 37
Conjugate functions 99
Continued fractions 211
Contractifiable semigroups 315
Contractions 115, 315
Control theory 115
Convariance matrix 149
Convex cones 99
Convexity 55

D

Derivatives of the Grüneisen and Einstein functions 175
Digraphs 235
Diophantine approximation 241
Diophantine approximation of roots of positive integers 241
Diophantine equation 241
Diophantine equations 31
Distribution of eigenvalues of covariance matrices of residuals in analysis of variance 149
Duality 99
Dynamic programming 245

E

Effective computability 241
Eigenvalues 149, 187
Einstein function 175
Eisenstein series 131
Electrical resistivity 175
Eoectronic energies 85
Element of finite order 231
Entire function 55
Entire functions of exponential type 55
Enumeration problem for a congruence equation 37
Error estimates 251
Error estimates for the solution of linear algebraic systems 251
Error function 211
Evaluation of computer programs 251
Expansions 155
Expansions with coefficient algorithms for time domain responses of skin effect lossy coaxial cables 155
Exponential type 55

F

F-group 231
Fitzgerald, K. E., Error estimates for the solution of linear algebraic systems 251
Functional analysis 115, 315

G

Gegenbauer functions 187
Geller, M., Ng, E. W., On some indefinite integrals of confluent hypergeometric functions 85
Generalized matrix function 45
Generating functions 1
Genus 121
Graphs 235
Greenberg, H. J., The minimum number of problems to cover all subproblems 245
Greenberg, L., Newman, M., Normal subgroups of the modular group 121
Gross, F., Entire functions of exponential type 55
Grüneisen function 175
Gupta, H., Partitions—A survey 1

H

Heat-conducting 67
Hecke basis theorems for groups of genus O 131
Hecke group 131
H-group 131
Hermitian matrix 45
Holt, R. D., Expansions with coefficient algorithms for time domain responses of skin effect lossy coaxial cables 155

I

Identities 1
Improper integral 183
Indefinite integrals 85
Index 121
Interaction 149
Inverse Laplace transform 155
Inverse of a matrix 251
Inviscid 67
Irreducible character 41
Isothermal 67

K

Kirchhoff-Langevin equation 67
Knopp, M. I., Smart, J. R., Hecke basis theorems for groups of genus O 131

L

Legendre functions 187
Lehner, J., A property of the triangle groups 231
Linear programming 99
Linear systems and matrices 251

M

Mandel, J., The distribution of eigenvalues of covariance matrices of residuals in analysis of variance 149
Mathieu functions 187
Matrices 311
Matrix 149
Maximum modulus 55
McKinney, J. E., Oser, H. J., Acoustic propagation and stability within an inviscid heat-conducting fluid 67
Merris, R., Partitioned hermitian matrices 45
Meyers, P. R., Contractifiable semigroups 315
———, On contractive semigroups and uniform asymptotic stability 115
Minimum number of problems to cover all subproblems 245
Modular form 131
Modular group 121

N

Newman, M., Brualdi, R. A., An enumeration problem for a congruence equation 37
———, *Greenberg, L.*, Normal subgroups of the modular group 121
Ng, E. W., Geller, M., On some indefinite integrals of confluent hypergeometric functions 85
Nonlinear transformation 183
Normal subgroup 231
Normal subgroups 121
Normal subgroups of the modular group 121

O

On the singular values of a product matrices ... 311
On complementary polar conical sets ... 99
On contractive semigroups and uniform asymptotic stability ... 115
On some indefinite integrals of confluent hypergeometric functions ... 85
On the spheroidal functions ... 187
Optimization ... 245
Orthogonal decomposition of tensor spaces ... 41
Orthogonality ... 99
Orthogonalization ... 187
Oser, H. J., McKinney, J. E., Acoustic propagation and stability within an inviscid heat-conducting fluid ... 67
Osgood, C. F., The diophantine approximation of roots of positive integers ... 241

P

Parametrization of cusp forms ... 131
Partitioned hermitian matrices ... 47
Partitions ... 1
Partitions—A survey ... 1
Permanents ... 37
Pierce, S., Orthogonal decomposition of tensor spaces ... 41
Polarity ... 99
Polynomials ... 47
Positive semi-definite hermitian matrix ... 45
Products ... 235
Property of the triangle groups ... 231
Prouhet-Tarry-Escott problem ... 31

R

Recurrence formulas ... 1
Recurrence relations ... 211
Reflectometry ... 155
Relation between the coefficients and roots of two equations and its application to diophantine problems ... 31
Residuals ... 149
Rhodes, D. R., On the spheroidal functions ... 187
Rivlin, T. J., Bounds on a polynomial ... 47

S

Selecting nonlinear transformations for the evaluation of improper intergrals ... 183
Semigroup ... 115
Series ... 211
Singular values ... 311
Sinha, T. N., A relation between the coefficients and roots of two equations and its application to diophantine problems ... 31
Skin effect ... 155
Smart, J. R., Knopp, M. I., Hecke basis theorems for groups of genus O ... 131
Solid-state physics ... 175
Some theorems on tensor composite graphs ... 235
Stability ... 67
Stability theory ... 115, 315
Steady-state ... 67
Stegun, I. A., Zucker, R., Automatic computing methods for special functions ... 211
Step response ... 155
Symmetric functions ... 31
Symmetry class of tensors ... 41
Symmetry operator ... 41
System function ... 155
Systems of distinct representatives ... 245

T

Tensor ... 235
Tensor product ... 41
Theory of computation ... 245
Thermal ... 67
Topology ... 115
Transient ... 67
Transmission ... 155
Transport properties ... 175
Trees ... 235
Triangle group ... 231
Two-way table ... 149

V

Vacuum cleaner ... 149

W

Watkins, W., On the singular values of a product matrices ... 311
Witzgal, C., On complementary polar conical sets ... 99

Z

Zucker, R., Stegun, I. A., Automatic computing methods for sperial functions ... 211

U.S. GOVERNMENT PRINTING OFFICE : 1971 O—406-680

UNITED STATES
GOVERNMENT PRINTING OFFICE
PUBLIC DOCUMENTS DEPARTMENT
WASHINGTON, D.C. 20402

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

U.S. GOVERNMENT PRINTING OFFICE
POSTAGE AND FEES PAID

JOURNAL OF RESEARCH of the National Bureau of Standards Vol. 74B, No. 4 Oct.–Dec. 1970
B. Mathematical Sciences

Contents

	Page
A property of the triangle groups. Joseph Lehner.	**231**
Some theorems on tensor composite graphs. Michael F. Capobianco.	**235**
The diophantine approximation of roots of positive integers. Charles F. Osgood.	**241**
The minimum number of problems to cover all subproblems. H. J. Greenberg.	**245**
Error estimates for the solution of linear algebraic systems. Brother Kenneth E. Fitzgerald, F.S.C.	**251**
On the singular values of a product matrices. William Watkins.	**311**
Contractifiable semigroups. Philip R. Meyers.	**315**
Publications of the National Bureau of Standards.	**323**
Index to Volume 74B, January to December 1970.	**331**

Library of Congress Catalog Card Number: 62–4414

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. (Order by SD Catalog No. C 13.22/sec.B:74/4). Price $1.25 (single copy). Subscription price: $5.00 a year; $1.25 additional for foreign mailing.
UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1970